UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
UNIDADE ACADÊMICA DE MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA

Benedito Vicente dos Santos

Um Estudo Comparativo entre Estatísticas *Scan* Espacial Baseadas em Modelos de Regressão com Variáveis Respostas Assimétricas

CAMPINA GRANDE - PB

2020

Universidade Federal de Campina Grande
Centro de Ciências e Tecnologia
Programa de Pós-Graduação em Matemática
Curso de Mestrado em Matemática

Um Estudo Comparativo entre Estatísticas *Scan* Espacial Baseadas em Modelos de Regressão com Variáveis Respostas Assimétricas

por

Benedito Vicente dos Santos [†]

sob orientação do

Prof. Dr. Manoel Ferreira dos Santos Neto

Dissertação apresentada ao Corpo Docente do Programa de Pós-Graduação em Matemática - CCT - UFCG, como requisito parcial para obtenção do título de Mestre em Matemática.

CAMPINA GRANDE - PB

2020

[†]Este trabalho contou com apoio financeiro da CAPES

Um Estudo Comparativo entre Estatísticas *Scan* Espacial Baseadas em Modelos de Regressão com Variáveis Respostas Assimétricas

por

Benedito Vicente dos Santos

Dissertação apresentada ao Corpo Docente do Programa de Pós-Graduação em Matemática - CCT - UFCG, como requisito parcial para obtenção do título de Mestre em Matemática.

Área de Concentração: Probabilidade e Estatística

Aprovada por:

Prof. Dr. João Batista Carvalho - UFCG

Prof. Dr. Max Sousa de Lima - UFAM

Prof. Dr. Manoel Ferreira dos Santos Neto - UFCG

Orientador

Universidade Federal de Campina Grande

Centro de Ciências e Tecnologia

Programa de Pós-Graduação em Matemática

Curso de Mestrado em Matemática

20 de agosto de 2020

S237e Santos, Benedito Vicente dos.
Um estudo comparativo entre estatísticas scan espacial baseadas em modelos de regressão com variáveis respostas assimétricas / Benedito Vicente dos Santos. - Campina Grande, 2020.
88 f. : il. : color.

Dissertação (Mestrado em Matemática) - Universidade Federal de Campina Grande, Centro de Ciências e Tecnologia, 2020.
"Orientação: Prof. Dr. Manoel Ferreira dos Santos Neto.
Referências.

1. Estatística Scan. 2. Respostas Assimétricas. 3. Simulações de Monte Carlo. I. Santos Neto, Manoel Ferreira dos. II. Título.

CDU 51(043)

FICHA CATALOGRÁFICA ELABORADA PELO BIBLIOTECÁRIO GUSTAVO DINIZ DO NASCIMENTO CRB-15/515

Resumo

As estatísticas *scan* espacial foram desenvolvidas para a detecção geográfica de *clusters* em diferentes tipos de modelos probabilísticos. Nesta dissertação, apresentamos algumas estatísticas *scan* espacial baseadas em alguns modelos de regressão com variável resposta assimétrica. As estatísticas de teste são baseadas em um teste de razão de verossimilhança e avaliadas usando o p-valor do *Bootstrap*. O poder estatístico, a sensibilidade e o valor preditivo positivo do teste são examinados através de um estudo de simulação de Monte Carlo. Por fim, uma aplicação a dados de tuberculose do estado do Maranhão é feita usando o modelo que apresentou o melhor comportamento nas simulações.

Palavras-chave: Estatística *Scan*, Respostas Assimétricas, Simulações de Monte Carlo.

Abstract

Spatial Scan Statistics has been developed as a geographical cluster detection analysis tool in different probabilistic models. In this dissertation, we present some spatial scan statistics based on some regression models with an asymmetric response variable. The test statistics are based on a likelihood ratio test and evaluated using Bootstrap p-value. The statistical power, sensitivity and positive predicted value of the test are examined through a Monte Carlo simulation study. Finally, an application to pulmonary tuberculosis data in state of Maranhão is made using the model that showed good results in the simulations.

Keywords: Scan statistics, Asymmetric responses, Monte Carlo simulations.

Agradecimentos

Aos meus pais, Aniceto e Lúcia, que sempre me apoiaram em todos os momentos.

Aos meus irmãos, Iris, Celson, Maísa, Cristina, Rafaela e Karl Marx pelo apoio e compreensão.

Ao professor Manoel Ferreira dos Santos Neto por me orientar na construção deste trabalho, pelo apoio, dedicação e compreensão em todos os momentos da realização desta dissertação.

À professora Michelli Karinne Barros da Silva, por sua ajuda e pelos conhecimentos transmitidos, muito obrigado.

Ao professor João Batista Carvalho, por aceitar o convite para participar da banca examinadora.

Ao professor Max Souza de Lima, por aceitar o convite para participar da banca examinadora, por fornecer o conjunto de dados e no auxílio na aplicação.

As colegas do mestrado, Raquel e Camila, pela boa convivência ao longo do curso.

Aos professores do Departamento de Estatística da UEPB, que contribuíram na minha formação acadêmica na graduação.

Aos professores Maria Joseane Cruz da Silva e Joelson da Cruz Campos por suas contribuições na minha formação acadêmica no mestrado.

À CAPES pelo apoio financeiro.

Dedicatória

Aos meus pais e irmãos.

Lista de Figuras

Lista de Tabelas

"Por aqui, contudo, não olhamos para trás por muito tempo.
Seguimos em frente, abrindo novas portas e fazendo coisas novas...
E a curiosidade nos conduz a novos caminhos."
(WALT DISNEY)

Sumário

Capítulo 1

Introdução

A estatística *scan* espacial (KULLDORFF, 1997) é um método amplamente utilizado em problemas de agrupamento espacial atualmente. Exemplos de uso dessa metodologia podem ser encontrados em muitas áreas diferentes, como detecção precoce de surtos de doenças (BESCULIDES et al., 2005), vigilância sindrômica (WIJNGAARD et al., 2010), criminologia (MINAMISAVA et al., 2009), doenças infecciosas (ELIAS et al., 2006), medicina (HUANG et al., 2010), educação básica e mortalidade infantil (LIMA et al., 2016) e emissões de bilhetes (CANÇADO; FERNANDES; SILVA, 2017).

A popularidade da estatística de Kulldorff deve-se a uma série de fatores, como a simplicidade computacional e a eficiência, fornecidos pelo processo inferencial de adaptação natural que governa as escolhas na detecção de *clusters* espaciais. A estatística *scan* espacial foi originalmente proposta para dados de contagem, particularmente para os modelos de Poisson e Bernoulli, mas extensões para ordinal (JUNG; KULLDORFF; KLASSEN, 2007), multinomial (JUNG; KULLDORFF; RICHARD, 2010), exponencial (HUANG; KULLDORFF; GREGORIO, 2007), normal (KULLDORFF; HUANG; KONTY, 2009), Weibull (BHATT; TIMARI, 2014) e beta (LIMA et al., 2016) também foram propostos.

Nesta dissertação, as estatísticas *scan* espacial propostas são baseadas em modelos de regressão com variáveis repostas assimétricas positivas. Os modelos de probabilidade frequentemente utilizados para modelagem de dados assimétricos positivos são os modelos normal inverso (NI), gama (GA), Weibull (WE), Birnbaum-Saunders (BS) e

beta-prime (BP). Situações que podem gerar dados assimétricos positivos são encontradas nas áreas de economia, análise de sobrevivência e médica, entre outras.

A distribuição NI é usada para descrever vários fenômenos e analisar quantitativamente os fenômenos em vários campos, não apenas em estatística matemática, mas também em engenharia (SATO; INOUE, 1994). Essa distribuição está intimamente relacionada à distribuição gaussiana (ou normal), conforme sugerido por seu nome. Além disso, é usado para descrever o primeiro tempo de passagem de uma partícula (movendo-se com velocidade constante) que está sujeita ao movimento browniano linear.

A distribuição GA incompleta, comumente chamada de distribuição GA, proposta por Thom (1947), é um caso especial da distribuição de Pearson tipo III, onde o parâmetro de posição é zero. Este modelo é frequentemente usado em meteorologia e climatologia. Adicionalmente, algumas distribuições bastante conhecidas são casos particulares da distribuição GA, como, por exemplo, as distribuições exponencial, qui-quadrado e Erlang.

O modelo probabilístico WE bi-paramétrico tem sido amplamente utilizado para análise de dados de sobrevivência em aplicações médicas e de engenharia. Por exemplo, esta distribuição é uma opção atraente para modelagem de sobrevivência totalmente paramétrica, uma vez que, unicamente, ela possui o tempo de falha acelerado e a propriedade de riscos proporcionais.

A distribuição BS (BIRNBAUM; SAUNDERS, 1969) tem recebido considerável atenção nos últimos anos, devido aos seus argumentos teóricos associados aos processos de danos cumulativos, suas propriedades e sua relação com a distribuição normal. Este modelo corresponde a uma distribuição unimodal, inclinada positivamente, de dois parâmetros e com suporte não negativo. Santos-Neto et al. (2012) propuseram várias parametrizações para a distribuição BS. Uma delas é indexada pela média e um parâmetro de precisão, que denotaremos por BS reparameterizada (BSR). É ela que será considerada no desenvolvido desta dissertação. Leiva et al. (2015), Leiva (2016), Santos-Neto et al. (2016) e Leão et al. (2017) mostraram que a distribuição BSR é útil em configurações para as quais a parametrização original apresenta algum tipo de limitação.

A distribuição BP (KEEPING, 1962; MCDONALD, 1984) também chamada de distribuição beta invertida ou distribuição beta do segundo tipo, é uma distribuição

alternativa de tempo de falha, como as distribuições BS, GA, NI e WE. Desta forma, é possível encontrar aplicações da distribuição BP em diferentes áreas do conhecimento, como por exemplo, na análise de dados de sobrevivência.

1.1 Objetivos

1.1.1 Geral

- Comparar estatísticas *scan* espacial baseadas em modelos de regressão com variáveis respostas assimétricas quanto ao desempenho em identificar *clusters* espaciais.

1.1.2 Específicos

- Propor modelos alternativos para estatísticas scan espacial;
- Apresentar os novos métodos de estatísticas scan espacial NI-SCAN, GA-SCAN, BSR-SCAN, BP-SCAN e WE-SCAN;
- Utilizar simulação de Monte Carlo (JOHANSEN et al., 2010) para gerar distribuições empíricas de estatísticas de teste;
- Utilizar o poder do teste, sensibilidade e valor preditivo positivo como medidas de desempenho dos modelos.

1.2 Organização da Dissertação

Além desta introdução a dissertação é composta por mais 5 Capítulos. No Capítulo 2, foi realizado uma breve descrição da estatística espacial e apresentaremos a estatística *scan* circular de Kulldorff, suas principais propriedades e conceitos. É também apresentada a estatística da razão de verossimilhanças e algumas medidas de desempenho.

No Capítulo 3, primeiro apresentaremos uma expressão geral do modelo de regressão com variável resposta assimétrica. Em seguida, apresentamos a expressão da estatística da razão de verossimilhanças para este modelo. Revisamos os modelos NI, GA, WE, BSR e BP e apresentamos as estatísticas *scan* espacial para cada modelo. Por

fim, de uma maneira genérica apresentaremos o cálculo do p-valor via *Bootstrap* para a estatística espacial baseada no modelo $\mathcal{A}$-SCAN.

No Capítulo 4, é feito um estudo de simulação para avaliar o desempenho das estatísticas SCAN espacial para os modelos NI-SCAN, GA-SCAN, WE-SCAN, BSR-SCAN e BP-SCAN. No Capítulo 5, é realizada a análise de um conjunto de dados sobre tuberculose do estado do Maranhão considerando o modelo SCAN que apresentou melhor resultado nas simulações. Todos os resultados serão apresentados através de Tabelas e Figuras. Por fim, comentários gerais sobre o estudo e possíveis pesquisas futuras a serem desenvolvidas a partir desta dissertação são apresentados no Capítulo 6.

1.3 Ferramentas Computacionais

Os códigos das simulações, as análises e gráficos foram feitos/escritos usando a linguagem de programação **R** (TEAM, 2020) disponível gratuitamente no endereço <www.r-project.org/>. Mais precisamente, os modelos de regressão considerados nesta dissertação foram ajustados usandos os pacotes **R**: **gamlss** (RIGBY; STASINOPOULOS, 2005), **RBS** (SANTOS-NETO, 2020) e **BPmodel** (BOURGUIGNON; SANTOS-NETO; CASTRO, 2018). As simulações realizadas durante o desenvolvimento deste trabalho foram feitas usando os recursos computacionais do Centro de Ciências Matemáticas Aplicadas à Indústria (CeMEAI), financiados pela FAPESP (proc. 2013/07375-0). Por fim, a dissertação foi escrita usando o LaTeX (GRATZER, 1996).

Capítulo 2

Preliminares

Este capítulo tem caráter predominantemente teórico. Em que foi feito uma breve descrição da estatística espacial e apresentamos a estatística *scan* circular de Kulldorff, seu conceito e suas principais propriedades. Além disso, são apresentadas a estatística de teste e as medidas de desempenho. Sendo assim, este capítulo tem uma importância fundamental para os capítulos subsequentes desta dissertação.

2.1 Estatística Espacial

A estatística espacial é a área da estatística que permite analisar a localização espacial de eventos. Isto é, a estatística espacial identificar, localizar, visualizar e analisar a ocorrência de padrões de fenômenos distribuídos no espaço utilizando métodos exploratórios e inferenciais específicos. Assim sendo, a estatística espacial tem notável importância na estatística e amplo uso teórico e prático. As principais aplicações da estatística espacial encontram-se nas áreas de saúde (mapeamento de doenças e epidemiologia espacial), ambiental (monitoramento de problemas ambientais), análise criminal, genética de populações, astronomia, entre outras áreas.

2.2 Tipos de Dados em Estatística Espacial

Os três tipos de dados mais utilizados na estatística espacial são classificados em dados de processos pontuais, dados de área e dados de superfícies aleatórias. A seguir,

são descritos cada um desses tipos de dados.

O tipo mais simples é o de processos pontuais, em que, são representados por um par (s_{1l}, s_{2l}) que corresponde a coordenada geográfica de ocorrência do evento de interesse de um dado mapa S dividido em L localizações s_l, para $l = 1, \ldots L$. Por exemplo, a localização de crimes e ocorrência de uma determinada doença.

Dados de área (ou agregados) são obtidos na situação em que não estão disponíveis as coordenadas de cada ocorrência do evento, mas apenas o número total de ocorrências em cada região, por exemplo, a ocorrência de Sars-CoV-19 (ou Covid-19) nos bairros da cidade de Campina Grande.

Os dados de superfície são obtidos, quando é feito medições em determinadas localizações do mapa, em que, cada elemento do conjunto de dados são representados por (s_{1l}, s_{2l}, s_{3l}) que indica a coordenada geográfica associada à medição feita naquela localização. Exemplos clássicos de dados de superfície, são medição de temperatura ou umidade em determinadas localizações.

Segundo Lima (2011) um processo pontual pode ser transformado em dados de área. Assim sendo, neste trabalho o enfoque será nos processos espaciais modelados por dados de áreas (ou agregados). Desta maneira, imagine que existe um processo estocástico $\boldsymbol{Y}(s) = \{Y(s_l) : l = 1, \ldots, L\}$ que é uma coleção de variáveis aleatórias, isto é, para cada $l = 1, \ldots, L$, $Y(s_l)$ denota a variável aleatória do processo em uma determinada área A_l, identificada por um ponto $s_l \in S = \{s_1, \ldots, s_L\}$ que representa o centro do polígono limitado por A_l.

2.3 Estatística *Scan* Espacial

Nesta seção, é realizado um estudo da teoria do método da estatística de varredura (do inglês *scan*) espacial proposto em Kulldorff (1997). Este método foi desenvolvido para a detecção geográfica de agrupamentos (do inglês *clusters*) espaciais em diferentes tipos de modelos, por exemplo, Bernoulli, multinomial, Poisson, Exponencial, Weibull, Normal, log-Normal e Beta. Um *cluster* espacial é definido como um subconjunto de regiões de um mapa em que a ocorrência de casos de um fenômeno de interesse é discrepante do restante do mapa, ou seja, um número anormalmente alto ou baixo de casos em alguma área. Neste trabalho, o interesse é detectar *clusters* em termos de

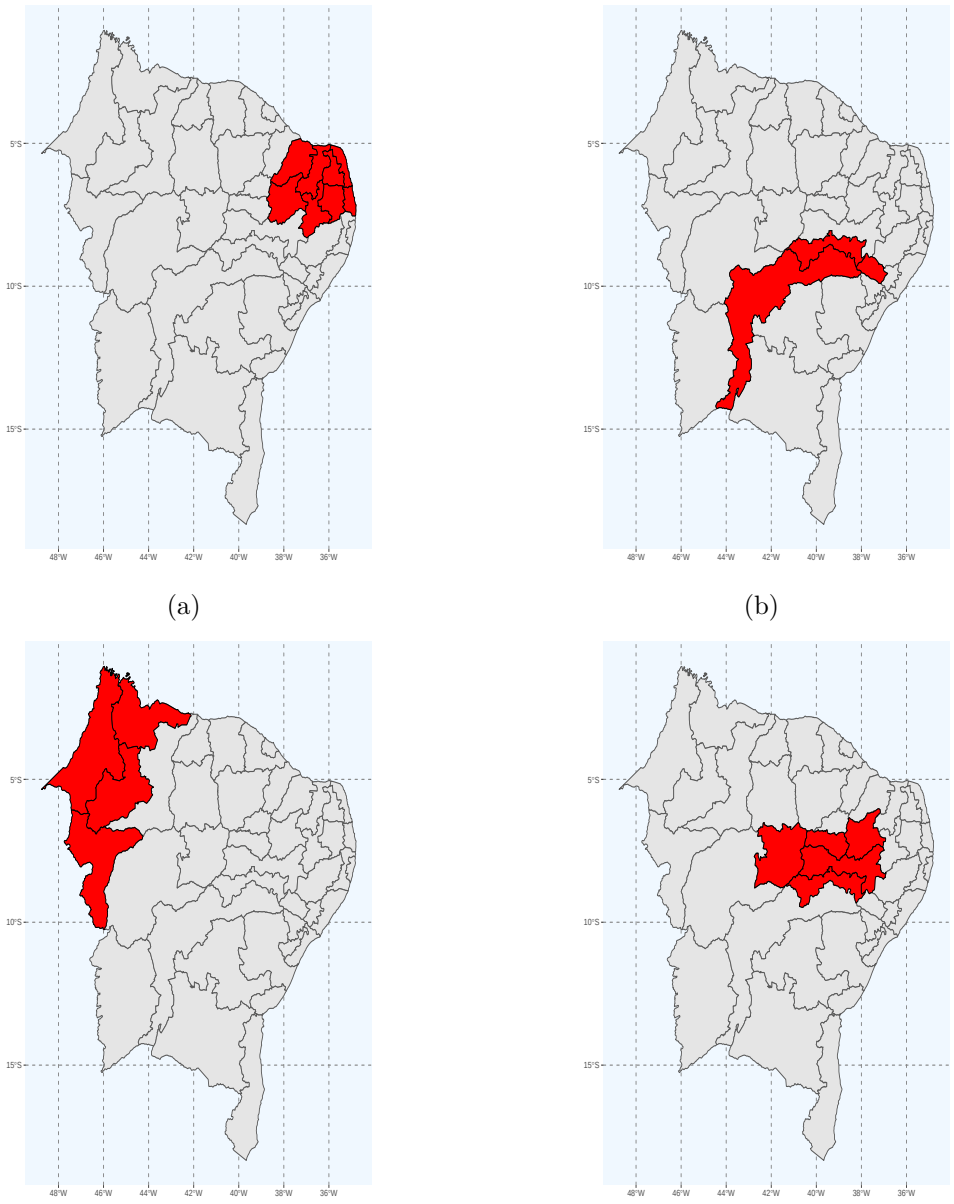
um número anormalmente alto de casos em alguma área. A detecção e localização dos *clusters* serão realizadas através de testes de hipóteses, em que o interesse é testar as seguintes hipóteses: H_0 não existe *cluster* no mapa versus H_1 existe *cluster* no mapa.

Além disso, a notação usada para os modelos de Bernoulli e Poisson nesta seção e também ao longo desta dissertação é a seguinte considere um mapa S dividido em L localizações (ou áreas ou regiões) $s_1, s_2, \ldots, s_L$, seja N a população total do mapa, Z é uma zona, ou seja, um subconjunto de regiões conexas do mapa a Figura 2.1 ilustra quatro zonas distintas dentro do mesmo mapa, $\mathcal{Z}$ conjunto das zonas Z, μ_Z é o número esperado de casos na zona Z, C é o número total de casos do mapa, p é a probabilidade de uma observação ser um caso dentro da zona Z, q é a probabilidade de ocorrer um caso fora da zona Z, c_Z é o número de casos observados na zona Z e n_Z a população observada na zona Z.

A estatística *scan* espacial (KULLDORFF, 1997) é um método bastante utilizado para a detecção e inferência de *clusters* espaciais. Este método utiliza uma janela de forma circular que se move através do mapa, também denominado região de estudo. Esta janela é centrada nos centróides das áreas avaliadas no mapa, sendo assim, os círculos construídos são capazes de incluir diferentes conjuntos de áreas, isto é, pode incluir áreas vizinhas. Observa-se que, uma área está incluída no círculo se o seu centróide está dentro da janela. Além disso, o tamanho da janela pode variar, ou seja, ela pode assumir qualquer tamanho conforme se move através do mapa. O raio de cobertura da janela para cada ponto onde o círculo é centrado, pode variar continuamente de zero até um valor máximo, além do mais, a janela nunca ultrapassa os 50% da população total do mapa incluído na janela circular. Deste modo, a janela circular é flexível em tamanho e localização. O raio da janela circular é limitado e será denotado por r. Portanto, o máximo de zonas circulares distintas que serão avaliadas pela estatística *scan* circular é L^2, que se torna computacionalmente viável. Neste método uma coleção $\mathcal{Z}$ de círculos distintos é formada, em que, cada diferente conjunto de áreas dentro de um determinado círculo é denominado de zona Z, e cada zona $Z \in \mathcal{Z}$ é considerado um provável candidato a *cluster*. A significância estatística da zona Z será avaliada por meio do teste da razão de verossimilhança.

A seguir vai ser apresentada a estatística *scan* circular espacial para os modelos de Bernoulli e Poisson.

Figura 2.1: Quatro diferentes zonas dentro de um mesmo mapa.



(a) (b)

(c) (d)

Fonte: Próprio autor

2.3.1 Modelo Bernoulli

De acordo com Kulldorff (1997), no modelo de Bernoulli há exatamente uma zona Z tal que cada observação dentro dessa zona tem probabilidade p de ser um caso, enquanto a probabilidade para cada observação fora da zona Z ser um caso é q. Nota-se que, essas probabilidades são independentes para todas as observações. Aqui cada observação corresponde a uma "entidade" ou "indivíduo" que pode está em qualquer

um de dois estados por exemplo presença ou ausência de uma doença em pessoas. Além disso, segundo Kulldorff (1997), observações em um desses estados são definidas como pontos, e a localização dessas observações constitui o processo pontual. Logo, a estatística *scan* de Kulldorff é definida a partir de um teste de razão de verossimilhança, em que o interesse é testar:

$$\begin{cases} \mathcal{H}_0: & p & = & q, \\ \mathcal{H}_1: & p & > & q, \quad Z \in \mathcal{Z}. \end{cases}$$

Nota-se que, sob $\mathcal{H}_0$, não existe *cluster* no mapa, ou seja, a probabilidade de ocorrência de um caso é a mesma em qualquer área do mapa. No caso em que o teste rejeitar $\mathcal{H}_0$, a zona Z será considerada um *cluster*. Portanto, a verossimilhança sob $\mathcal{H}_0$ é dada por:

$$L_0(p) = p^C(1-p)^{N-C},$$

segue que o logaritmo da função de verossimilhança é dado por:

$$\begin{aligned} \ell_0(p) =& \log(L_0(p)) \\ =& \log[p^C(1-p)^{N-C}] \\ =& C\log(p) + (N-C)\log(1-p). \end{aligned}$$

Derivando em relação a p e igualando a zero, isto é:

$$\frac{\partial \ell_0(p)}{\partial p}\Big|_{p=\hat{p}} = \frac{C}{\hat{p}} - \frac{(N-C)}{1-\hat{p}} = 0,$$

segue que,

$$\begin{aligned} \frac{C(1-\hat{p}) - \hat{p}(N-C)}{\hat{p}(1-\hat{p})} =& 0 \\ C - C\hat{p} - \hat{p}N + \hat{p}C =& 0 \\ C =& \hat{p}N \\ \hat{p} =& \frac{C}{N}. \end{aligned}$$

Sob $\mathcal{H}_1$, a função de verossimilhança será escrita como:

$$L(Z,p,q) = p^{c_Z}(1-p)^{n_Z-c_Z}q^{C-c_Z}(1-q)^{(N-C)-(n_Z-c_Z)},$$

o logaritmo da função de verossimilhança é dado por:

$$\begin{aligned}\ell(Z,p,q) =& \log(L(Z,p,q))\\ =& \log[p^{c_Z}(1-p)^{n_Z-c_Z}q^{C-c_Z}(1-q)^{(N-C)-(n_Z-c_Z)}]\\ =& c_Z\log(p) + (n_Z-c_Z)\log(1-p) + (C-c_Z)\log(q)\\ &+ [(N-C)-(n_Z-c_Z)]\log(1-q).\end{aligned}$$

Os estimadores de máxima verossimilhança são soluções das equações abaixo:

$$\frac{\partial \ell(Z,p,q)}{\partial p}\Big|_{p=\hat{p}} = 0, \qquad \frac{\partial \ell(Z,p,q)}{\partial q}\Big|_{q=\hat{q}} = 0,$$

em que obtemos,

$$\hat{p} = \frac{c_Z}{n_Z}, \qquad \hat{q} = \frac{C-c_Z}{N-n_Z}.$$

Com o objetivo de detectar a zona mais provável de ser um *cluster*, dentre todas as possíveis zonas candidatas, utiliza-se o teste da razão de verossimilhança, em que para o modelo de Bernoulli este teste é dado por:

$$\begin{aligned}\lambda =& \frac{\sup_{Z\in\mathcal{Z},p>q}L(Z,p,q)}{\sup_{p=q}L(Z,p,q)} = \frac{L(\hat{Z})}{L_0},\\ =& \frac{\left(\frac{c_Z}{n_Z}\right)^{c_Z}\left(1-\frac{c_Z}{n_Z}\right)^{n_Z-c_Z}\left(\frac{C-c_Z}{N-n_Z}\right)^{C-c_Z}\left(1-\frac{C-c_Z}{N-n_Z}\right)^{(N-C)-(n_Z-c_Z)}}{\left(\frac{C}{N}\right)^{C}\left(1-\frac{C}{N}\right)^{N-C}}.\end{aligned}$$

Além do mais, nota-se que o denominador de λ depende apenas do número total de casos e da população total do mapa e não depende da distribuição espacial das observações. Como não temos a expressão fechada da distribuição de λ sob $\mathcal{H}_0$, a mesma é estimada via simulações de Monte Carlo.

2.3.2 Modelo Poisson

No modelo de Poisson, segundo Kulldorff (1997), os casos são gerados por um processo de Poisson não homogêneo. Além disso, considere que no modelo, há exatamente uma zona Z. As hipóteses de interesse a serem contrastadas são formuladas da seguinte maneira:

$$\begin{cases} \mathcal{H}_0: & p = q, \\ \mathcal{H}_1: & p > q, \quad Z \in \mathcal{Z}. \end{cases}$$

Observa-se que, sob $\mathcal{H}_0$, a probabilidade de que uma observação seja um caso é a mesma em qualquer área do mapa. Então de acordo com a hipótese $\mathcal{H}_0$, não existe *cluster* no mapa. Por outro lado, a hipótese $\mathcal{H}_1$ é de que existe uma zona $Z \in \mathcal{Z}$ que é um *cluster*. Considere $\mu_{\bar{Z}}$ como sendo o número de casos esperados fora da zona Z, então sob $\mathcal{H}_0$ tem-se $\mu_Z = pn_Z$ e $\mu_{\bar{Z}} = p(N - n_Z)$. Logo, a verossimilhança sob $\mathcal{H}_0$ é dada por:

$$L_0(Z,p) = \frac{(pn_Z)^{c_Z} e^{-pn_Z}}{c_Z!} \times \frac{[p(N-n_Z)]^{C-c_Z} e^{-p(N-n_Z)}}{(C-c_Z)!}.$$

O logaritmo da função de verossimilhança é dado por:

$$\begin{aligned} \ell_0(Z,p) =& \log(L_0(Z,p)) \\ =& \log\left[\frac{(pn_Z)^{c_Z} e^{-pn_Z}}{c_Z!} \times \frac{[p(N-n_Z)]^{C-c_Z} e^{-p(N-n_Z)}}{(C-c_Z)!}\right] \\ =& c_Z \log(p) + c_Z \log(n_Z) - pn_Z - \log(c_Z!) + (C-c_Z)\log(p) \\ &+ (C-c_Z)\log(N-n_Z) - p(N-n_Z) - \log((C-c_Z)!). \end{aligned}$$

Derivando em relação a p e igualando a zero isto é:

$$\frac{\partial \ell_0(Z,p)}{\partial p}\Big|_{p=\hat{p}} = 0 \implies \hat{p} = \frac{C}{N}.$$

Para a hipótese $\mathcal{H}_1$ defini-se $\mu_Z = pn_Z$ e $\mu_{\bar{Z}} = q(N-n_Z)$. Dessa forma, a verossimilhança sob $\mathcal{H}_1$ será escrita como:

$$L(Z,p,q) = \frac{(pn_Z)^{c_Z} e^{-pn_Z}}{c_Z!} \times \frac{[q(N-n_Z)]^{C-c_Z} e^{-q(N-n_Z)}}{(C-c_Z)!}.$$

O logaritmo da função de verossimilhança pode ser escrito da seguinte maneira:

$$\begin{aligned}\ell(Z,p,q) =& \log(L(Z,p,q)) \\ =& \log\left[\frac{(pn_Z)^{c_Z}e^{-pn_Z}}{c_Z!} \times \frac{[q(N-n_Z)]^{C-c_Z}e^{-q(N-n_Z)}}{(C-c_Z)!}\right] \\ =& c_Z\log(p) + c_Z\log(n_Z) - pn_Z - \log(c_Z!) + (C-c_Z)\log(q) \\ &+ (C-c_Z)\log(N-n_Z) - q(N-n_Z) - \log((C-c_Z)!).\end{aligned}$$

Os estimadores de máxima verossimilhança são soluções das equações abaixo:

$$\frac{\partial\ell(Z,p,q)}{\partial p}\Big|_{p=\hat{p}} = 0, \qquad \frac{\partial\ell(Z,p,q)}{\partial q}\Big|_{q=\hat{q}} = 0,$$

em que obtemos,

$$\hat{p} = \frac{c_Z}{n_Z}, \qquad \hat{q} = \frac{C-c_Z}{N-n_Z}.$$

O teste da razão de verossimilhança para o modelo de Poisson é dado por:

$$\lambda = \frac{\sup_{Z\in\mathcal{Z},p>q}L(Z,p,q)}{\sup_{p=q}L(Z,p,q)} = \frac{L(\hat{Z})}{L_0}.$$

$$\lambda = \begin{cases} \left(\dfrac{c_Z}{\mu_Z}\right)^{c_Z}\left(\dfrac{C-c_Z}{C-\mu_Z}\right)^{C-c_Z}, & \text{se} \quad \dfrac{c_Z}{\mu_Z} > \dfrac{C-c_Z}{C-\mu_Z}; \\ 1, & \text{caso contrário.} \end{cases}$$

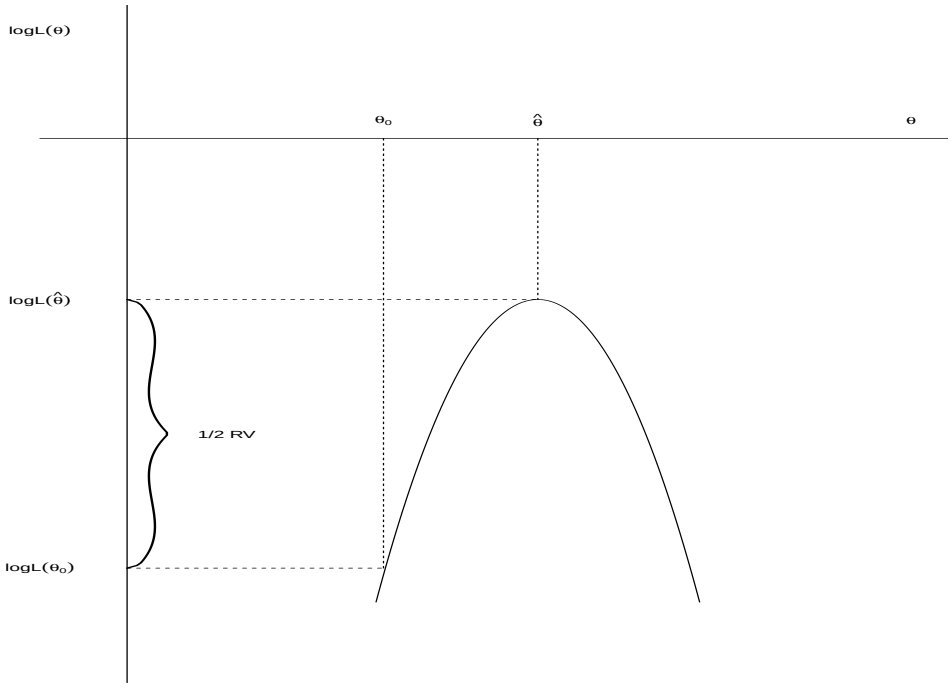
Como não temos a expressão fechada da distribuição de λ sob $\mathcal{H}_0$, a mesma é estimada via simulações de Monte Carlo.

Para escolher entre os modelos de Bernoulli e Poisson qual deve ser utilizado dependerá da aplicação. Por exemplo, para dados binários, como dois tipos de estrelas o modelo Bernoulli é o indicado. Para dados em forma de contagens de eventos deve ser usado o modelo de Poisson. De acordo com Kulldorff (1997) a escolha não importa muito quando o número total de casos é pequeno comparado a N, os modelos Bernoulli e Poisson aproximam-se um do outro.

2.4 Estatística de Teste

As estatísticas de teste apresentadas neste trabalho são baseadas no teste da razão de verossimilhança (BUSE, 1982). Uma descrição intuitiva desse teste pode ser vista na Figura 2.2.

Figura 2.2: Ilustração do teste da razão de verossimilhança.



Fonte: Próprio autor (Baseado em Buse (1982))

Suponha que as hipóteses de interesse a serem contrastadas são:

$$\begin{cases} \mathcal{H}_0: \ \theta \in \Omega_0, \\ \mathcal{H}_1: \ \theta \in \Omega_1, \end{cases}$$

com $\Omega_0 \cup \Omega_1 = \Omega$ e $\Omega_0 \cap \Omega_1 = \phi$, em que Ω e ϕ representam o espaço paramétrico e o conjunto vazio respectivamente, θ é o parâmetro desconhecido associado a um modelo de probabilidade de interesse podendo ser um escalar ou um vetor. A estatística de teste é baseada no teste da razão de verossimilhança, que pode ser escrito da seguinte forma:

$$RV = 2\{\ell(\hat{\theta}) - \ell(\tilde{\theta})\}.$$

em que, $\ell(\theta)$ representa à função de log-verossimilhança dos dados avaliada em θ, $\hat{\theta}$ é o estimador de máxima verossimilhança (EMV) de $\theta \in \Omega$, $\tilde{\theta}$ é o EMV de θ restrito à

hipótese nula ($\theta \in \Omega_0$).

Agora, considere o caso em que o parâmetro de interesse é um escalar e o objetivo é testar $\mathcal{H}_0 : \theta = \theta_0$ versus a hipótese alternativa $\mathcal{H}_1 : \theta \neq \theta_0$. Sendo assim, a estatística RV é escrita como $RV = 2\{\ell(\hat{\theta}) - \ell(\theta_0)\}$, nota-se na Figura 2.2, que a distância entre as log-verossimilhanças, $1/2RV$, dependerá da distância $\hat{\theta} - \theta_0$ bem como, da curva da função de log-verossimilhança, ou seja, para uma curva fixa, quanto maior a distância entre $\hat{\theta}$ e θ_0, maior será o valor de RV. Tem-se também que, para uma distância entre $\hat{\theta}$ e θ_0 fixa, quanto maior o valor da curva da função de log-verossimilhança maior o valor da estatística RV.

Mais formalmente no contexto da estatística espacial, considere uma variável aleatória de interesse $Y(s_l)$ definida na localização s_l com função densidade de probabilidade (f.d.p) ou função massa de probabilidade (f.m.p) denotada por $f(y_l; \theta)$, em que $Y(s_l)$ segue distribuição P_0 quando a probabilidade de ocorrência de um caso é a mesma em qualquer área do mapa S, ou seja, não existe cluster no mapa. Caso contrário, $Y(s_l)$ assumirá distribuição P_1. Logo, o interesse é testar as seguintes hipóteses:

$$\begin{cases} \mathcal{H}_0 : & Y(s_l) & \sim & P_0, & \forall s_l \in S, \\ \mathcal{H}_1 : & Y(s_l) & \sim & P_1, & \forall s_l \in Z. \end{cases}$$

A função de verossimilhança pode ser escrita da seguinte maneira:

$$L(\theta) = \prod_{l=1}^{L} f(y_l; \theta),$$

em que $\theta \in \Theta$ é uma quantidade desconhecida que denominamos de parâmetro do modelo $P(\cdot)$ e Θ representa todo o espaço paramétrico.

Considere que $Z \in \mathcal{Z}$ é uma zona. Seja $L(Z)$ a função de verossimilhança sob a hipótese alternativa $\mathcal{H}_1$, em que sob esta hipótese existe uma zona Z^* que é um *cluster*, e $L(0)$ a função de verossimilhança sob a hipótese nula $\mathcal{H}_0$ de que não existe um *cluster* no mapa, desse modo:

$$L(\theta) = \prod_{s_l \in \mathcal{Z}} f(y_l; \theta) \prod_{s_l \notin \mathcal{Z}} f(y_l; \theta).$$

Com o objetivo de identificar a zona mais verossímil de ser o *cluster* Z^*, dentre todas as possíveis zonas candidatas a *cluster*, usa-se o teste da razão de verossimilhança

que é dado por:

$$\Lambda^*(Z) = \frac{\sup_{\mathcal{H}_1} L(Z)}{\sup_{\mathcal{H}_0} L(0)}.$$

Assim, a zona Z mais provável será aquela que maximiza a função $\Lambda^*(Z)$ dentre todas as zonas candidatas a *cluster*. Nota-se que, a estatística de teste é definida por $\Lambda = \max_{Z \in \mathcal{Z}} \Lambda(Z)$. Além disso, da teoria estatística sabe-se que, a função $\Lambda(Z)$ assume valores muito grandes. Sendo assim, para suavizar esse problema, utiliza-se o logaritmo da razão de verossimilhança para $\Lambda(Z)$. Pois, como a função logaritmo é monotonicamente crescente e $\Lambda(Z)$ cresce muito rápido, maximizar $\log(\Lambda(Z))$ é equivalente a maximizar $\Lambda(Z)$. Portanto,

$$\Lambda(Z) = (\ell_Z - \ell_0).$$

Note que, uma vez identificada a zona mais provável, é necessário também verificar a sua significância estatística para que a zona detectada seja considerada um *cluster*. Como a distribuição exata da estatística de teste Λ é desconhecida, ou seja, é analiticamente intratável sobre $\mathcal{H}_0$. Deste modo, a avaliação da significância estatística da zona mais provável identificada nos dados observados no estudo é realizada por meio de simulação de Monte Carlo, segundo o procedimento proposto em Dwass (1957). Esse procedimento é baseado na geração de réplicas do mapa original, primeiramente casos simulados aleatoriamente são distribuídos sob $\mathcal{H}_0$ na região em estudo e a estatística do teste da razão de verossimilhança (estatística de teste) é calculada. Em seguida este procedimento é repetido milhares de vezes, ou seja, uma grande quantidade de vezes, em que o principal objetivo é obter uma distribuição empírica para a estatística de teste Λ, sob $\mathcal{H}_0$. Por fim, se compara o valor da estatística de teste dos dados observados no estudo com a distribuição empírica obtida na simulação para determinar seu nível descritivo (p-valor).

2.5 Representação Espacial dos *Clusters*

Para um mapa S dividido em L localizações $s_1, s_2, \ldots, s_L$, ou seja, $S = \{s_1, \ldots, s_L\}$, considere um centroide com coordenada geográfica $s_l = (s_{1l}, s_{2l})$ para a l-ésima área do mapa. Desta maneira, a distância ente duas localizações será calculada através da

distância euclidiana entre seus centroides. Formalmente, a distância entre dois centroides quaisquer é dada por:

$$d(l,m) = \sqrt{(s_{1l} - s_{1m})^2 + (s_{2l} - s_{2m})^2}.$$

A matriz de distâncias entre os centroides é simétrica e quadrada de ordem $L \times L$, com entrada na l-ésima linha e m-ésima coluna, denotada por $d(l,m)$ que corresponde a distância entre o par (l,m) de localizações. Observa-se que todas as entradas da matriz são não negativas e que cada elemento da matriz representa a distância entre duas localizações do mapa S. Logo, a matriz é definida por:

$$D = \begin{bmatrix} 0 & d_{1,2} & \dots & d_{1,m} & \dots & d_{1,L} \\ d_{21} & 0 & \dots & d_{2,m} & \dots & d_{2,L} \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ d_{l,1} & d_{l,2} & \dots & 0 & \dots & d_{l,L} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{L,1} & d_{L,2} & \dots & d_{L,m} & \dots & 0 \end{bmatrix}.$$

A partir daí, a próxima etapa será ordenar em ordem crescente os elementos da matriz D, guardando o índice l do centroide s_l e também o índice c do centroide mais próximo s_c respectivamente, em uma matriz de adjacência $\mathbb{I}$. Isto é:

$$\mathbb{I}_{l,m} = \begin{cases} l, & \text{se} \quad m = 1; \\ c, & \text{se} \quad s_c \text{ é o } m\text{-ésimo centroide mais próximo de } s_l. \end{cases}$$

Será mostrado a seguir um exemplo ilustrativo deste procedimento. Suponhamos um mapa com $L = 6$. Para a primeira linha, considere o vetor $(1,6,3,5,2,4)$. Desta forma, o centroide s_6 é o segundo mais próximo do centroide s_1, o centroide s_3 é o terceiro mais próximo do centroide s_1, o centroide s_5 é o quarto mais próximo do centroide s_1, o centroide s_2 é o quinto mais próximo do centroide s_1 e o centroide s_4 é o sexto mais próximo do centroide s_1. Sendo assim, deixando o centroide s_l fixo, podemos encontrar um provável *cluster* como sendo um vetor $Z_{l_i} = (l_{[i,1]}, l_{[i,2]}, \dots, l_{[i,L]})$ formado da seguinte maneira:

(i) Seja $l_{[i,m]} = 1,$ quando $l = m, i = 1,2,\dots,L$ e $m = 1,2,\dots,L$;

(ii) Seja $l_{[i,\mathbb{I}_{l,m}]} = 1$, quando $s_{\mathbb{I}_{l,m}}$ é um dos m centroides mais próximo de s_l e $m \leq i$. Caso contrário, $l_{[i,\mathbb{I}_{l,m}]} = 0$.

Por exemplo, considere um mapa $S = \{s_1, s_2, s_3, s_4, s_5, s_6\}$. Para a primeira linha, com a matriz de distância ordenada, suponhamos que obtemos o vetor $\{s_1, s_6, s_3, s_5, s_2, s_4\}$ de centroides. Assim, para Z_{l_i} tem-se $Z_{1_1} = (1, 0, 0, 0, 0, 0)$, $Z_{1_2} = (1, 0, 0, 0, 0, 1)$, $Z_{1_3} = (1, 0, 1, 0, 0, 1)$, $Z_{1_4} = (1, 0, 1, 0, 1, 1)$, $Z_{1_5} = (1, 1, 1, 0, 1, 1)$ e $Z_{1_6} = (1, 1, 1, 1, 1, 1)$. Nota-se que, para cada valor de m, Z_{l_i} recebe o valor 1 no índice do vizinho mais próximo do centroide s_l na sua posição original no espaço geográfico (mapa). Além disso, esta representação é única. É importante descrevemos como foi obtido esta representação dos candidatos a *cluters* acima, ou seja, por exemplo, para a formação da representação de Z_{l_2}, seja $l = 1, i = 2$ e $m = 1, 2, \ldots, L$.

(i) Quando $m = 1, l = j$ então, $1_{[1,1]} = 1$;

(ii) Quando $m = 2, \mathbb{I}_{1,2} = 6$, e s_6 é o segundo centroide mais próximo de s_1, observa-se que, $m = 2 \leq 2 = i$. Consequentemente, $1_{[1,6]} = 1$. No entanto, para todo $j \geq 3$ nesta situação $j > i$, logo não satisfaz a condição de que $j \leq i$ implicando que as outras coordenadas de Z_{1_2} sejam nulas. Portanto, $Z_{1_2} = (1, 0, 0, 0, 0, 1)$.

Repetindo o processo acima para as L áreas do mapa até atingir no máximo 50% da população total do mapa dentro da janela circular. Desta maneira, obtém-se todos os possíveis candidatos a *cluster*, $\tilde{Z} = \{Z_{l_i} : l, i = 1, 2, \ldots, L\}$. Adicionalmente, o número total de possíveis candidatos a *clusters* em $\tilde{Z}$ será L^2 e o raio máximo do círculo Z_{l_i} é dado por:

$$r_{l_i} = \max_{s_i \in Z_{l_i}} \{D_{l,i}\},$$

em que $D_{l,i}$ representa a distancia euclidiana. Quando obtidas todas as possíveis zonas candidatas a *clusters* $Z_{l_i} \in \mathcal{Z}$, o próximo passo será calcular a estatística de teste $\Lambda(Z_{l_i})$, ou seja,

$$\Lambda = \max\{\Lambda(Z_{l_i}) : i = 1, 2, \ldots, a; l = 1, 2, \ldots, L\},$$

com $\hat{Z} = \text{argmax}(\hat{\Lambda}(Z_{l_i}))$, e a é um valor fixo previamente definido, que indica a restrição do tamanho do *cluster* espacial.

2.6 Algoritmo *Scan* Circular

Nesta seção, iremos apresentar o algoritmo *scan* circular proposto por Kulldorff (1997). É importante observar que este algoritmo tem sido muito utilizado para detecção de *cluster* espacial por causa da sua simplicidade computacional e também por sua fácil implementação computacional. Este método utiliza uma janela de forma circular que se move através do mapa. Além disso, o tamanho da janela pode variar, ou seja, ela pode assumir qualquer tamanho conforme se move através do mapa. O raio de cobertura da janela para cada ponto onde o círculo é centrado, pode variar continuamente de zero até um valor máximo, em que, a janela nunca excede dos 50% da população total do mapa contido na janela. Sendo assim, a janela circular é flexível em tamanho e localização.

O algoritmo *scan* circular busca em todas as L^2 zonas possíveis a de maior valor de Λ_Z que define o *cluster* mais provável. Daí segue que para cada janela é calculada a razão de verossimilhança baseada no número esperado de eventos dentro e fora da janela, ou seja, a razão de verossimilhança permite em seu calculo ser interpretada como uma razão de chance, isto é, ela equilibra o quanto é grande o "risco" dentro e fora da janela em comparação ao "risco" calculado quando se considerar a hipótese nula verdadeira. O termo "risco" é comumente usado em epidemiologia para indicar a probabilidade de uma pessoa contrair uma determinada doença. Após calcular a razão de verossimilhança para todas as zonas, a significância da estatística do teste é realizada via simulação de Monte Carlo, em que, casos simulados aleatoriamente são distribuídos sob H_0 no mapa. As hipóteses definidas para este método são H_0 não existe *cluster* no mapa versus H_1 existe *cluster* no mapa. O algoritmo *scan* circular segue as seguintes etapas:

Etapa 1. Escolher uma região s_l no mapa S em estudo;

Etapa 2. Calcular as distâncias até as outras regiões (determinar a matriz D), ordenando-as em ordem crescente, e armazenando em um vetor (Z_{l_i});

Etapa 3. Criar um círculo centrado na região escolhida na etapa 1 e continuamente aumentar o seu raio conforme as distâncias encontradas na etapa 2. Para cada região s_l que entrar no círculo, atualizar $Y(s_l)$ dentro do círculo $\mathcal{Z}$. Calcular Λ_Z

para cada $Y(s_l)$. O *cluster* mais provável será aquele correspondente ao maior valor de Λ_Z;

Etapa 4. Repita as etapas 1, 2 e 3 para cada região do mapa S;

Etapa 5. Utilizar simulações de Monte Carlo para avaliar a significância da estatística do teste;

Etapa 6. Se a hipótese nula for rejeitada, ao nível de significância de 5%, então a zona $\hat{Z}$ associada com a maximização de Λ_Z é o *cluster* mais admissível e deve ser armazenada para que se faça o mapa destacando o *cluster* encontrado.

2.7 Propriedades da Estatística *Scan* Circular

As principais propriedades, da estatística *scan* circular proposta por Kulldorff serão descritas nesta seção. Primeiramente, serão apresentadas as vantagens desse método em comparação aos demais métodos de detecção de *cluster* espacial existentes na literatura. As suas principais vantagens são:

(i) Pode ser ajustada para a densidade populacional não constante no mapa.

(ii) Este método pode ser ajustado por variáveis de confusão, por exemplo idade e sexo.

(iii) O teste estatístico de razão de verossimilhança leva em conta testagens múltiplas informando um único p-valor ao testar a hipótese nula (KULLDORFF, 1997).

(iv) O método procura *clusters* sem especificar seu tamanho e localização.

(v) O método tem a capacidade de detectar a localização de *clusters* e também fazer inferência ao mesmo tempo. No sentido que, quando a hipótese nula é rejeitada, o método indica a localização do *cluster* mais verossímil que provocou à rejeição.

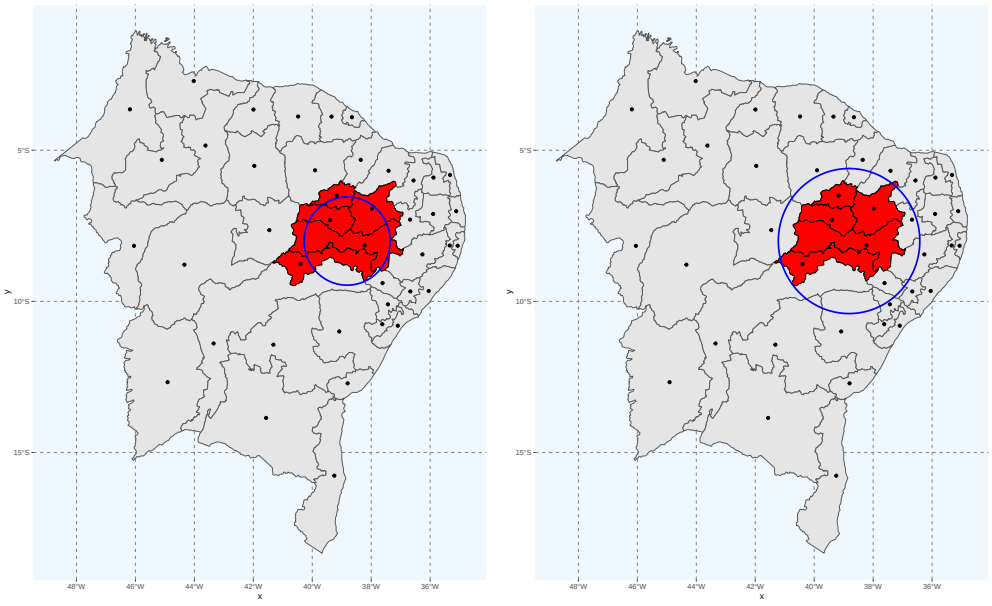
(vi) O método é uniformemente mais poderoso. De acordo com Casella e Berger (2002), um teste uniformemente mais poderoso é um teste de hipótese que tem o maior poder (probabilidade do teste rejeitar corretamente a hipótese nula) entre todos os possíveis testes de um dado tamanho.

Uma das principais desvantagens da estatística *scan* circular de Kulldorff em comparação aos demais métodos de detecção de *cluster* espacial existentes na literatura é a seguinte:

(i) O método fixa a forma geométrica dos candidatos a *clusters* como círculos. Ou seja, a principal desvantagem apresentada pelo método consiste no fato de que ele não tem a capacidade de detectar corretamente um *cluster* com forma muito diferente da circular.

Além disso, a estatística *scan* espacial de Kulldorff é mais adequada para detecção de *cluster* quando há apenas um único *cluster* bem definido na hipótese alternativa, pois segundo Kulldorff (1997) nesta situação a estatística *scan* de Kulldorff apresenta grande poder de teste, ou seja, o teste é uniformemente mais poderoso para detecção de *clusters*. No entanto, para situações em que o mapa apresenta mais de um *cluster* ou *cluster* de formato muito diferente do circular o poder do teste diminui como observado por, Kulldorff, Tango e Park (2003) e Duczmal, Kulldorff e Huang (2006). A redução do poder do teste é a principal causa da subestimação (*cluster* detectado menor do que o *cluster* real), ou à superestimação (*cluster* detectado maior do que o *cluster* real), como ilustrado na Figura 2.3.

Figura 2.3: Subestimação do *cluster* em (a). Superestimação do *cluster* em (b).



(a) (b)

Fonte: Próprio autor

2.8 Medidas de Desempenho

Para avaliar o desempenho das estatísticas *SCAN* espacial baseadas nos modelos propostos nesta pesquisa, que são os modelos NI-SCAN, GA-SCAN, WE-SCAN, BSR-SCAN e BP-SCAN, utilizou-se as medidas de sensibilidade, valor preditivo positivo e poder do teste, obtidas por meio de simulações de Monte Carlo.

O poder de detecção do *cluster* depende dos dados gerados nas simulações realizadas no estudo e também dependerá dos diferentes valores dos parâmetros de regressão, precisão e razão que compõem o modelo. Dessa forma, o poder do teste é definido como a probabilidade do teste detectar um *cluster* quando este realmente existe. Então o poder é calculado através da proporção de vezes em que H_0 foi rejeitada ao longo das simulações realizadas. A Sensibilidade (SS) e o Valor Preditivo Positivo (VPP) são definidos a seguir.

Sensibilidade (SS):quociente médio da população em risco corretamente detectada pela população em risco no *cluster* verdadeiro:

$$\mathrm{SS} = \frac{1}{N}\sum_{q=1}^{N}\left(\frac{\mathrm{pop}\{\widehat{Z}^{(q)} \cap Z\}}{\mathrm{pop}\{Z\}}\right),$$

Valor Preditivo Positivo (VPP): o quociente médio da população em risco no *cluster* verdadeiro pela população em risco corretamente detectada:

$$\mathrm{VPP} = \frac{1}{N}\sum_{q=1}^{N}\left(\frac{\mathtt{pop}\{\widehat{Z}^{(q)} \cap Z\}}{\mathrm{pop}\{\widehat{Z}^{(q)}\}}\right),$$

em que, N é o total de simulações no estudo, $\widehat{Z}^{(q)}$ é o *cluster* detectado na q-ésima simulação, Z é o *cluster* verdadeiro no cenário correspondente e $\mathrm{pop}\{A\}$ é o conjunto de indivíduos em risco da coleção de áreas A. Portanto, SS e VPP avaliam a capacidade do método de localizar o *cluster*, quando ele existe.

Nota-se que, a sensibilidade indica o quanto do *cluster* verdadeiro é detectado, por outro lado, o VPP representa o quanto do *cluster* detectado pertence ao verdadeiro. O valor de SS e VPP pertence ao intervalo $(0, 1)$, além disso, quanto mais próximo de 1 os valores dessas duas medidas juntas, indica alta precisão para detectar o local correto do *cluster*. No entanto, um grande valor de VPP e pequeno valor de SS juntos não mostra uma boa precisão e vice-versa.

Capítulo 3

Estatísticas *Scan* Espacial Baseadas em Modelos de Regressão com Variáveis Respostas Assimétricas

Este capítulo também é de caráter predominantemente teórico. De início, apresentamos o modelo de regressão com variável resposta assimétrica e a estatística de teste para esse modelo. Em seguida, revisamos os modelos NI, GA, WE, BSR e BP e apresentamos as estatísticas *scan* espacial para cada modelo. Por fim, apresentamos o cálculo do p-valor via *Bootstrap* para a estatística espacial $\mathcal{A}$-SCAN.

3.1 Modelo de Regressão com Resposta Assimétrica

Dado L localização espacial s_l, seja $\boldsymbol{Y} = [Y(s_1), \ldots, Y(s_L)]^\top$, onde $Y_l \equiv Y(s_l)$ é uma variável aleatória no intervalo $(0, \infty)$. Mais especificamente, assuma que Y_l segue uma distribuição assimétrica $Y_l \sim \mathcal{A}(\mu_l, \sigma)$, em que μ_l denota o parâmetro de locação e σ o parâmetro de dispersão. O modelo de regressão $\mathcal{A}$-SCAN é definido da seguinte forma. Seja Z um potencial *cluster*, seguindo o processo espacial $Y_1, \ldots, Y_L$. Este processo é modelado por $\mathcal{A}$-SCAN$(\mu_l, \sigma, \tau), l = 1, \ldots, L$ quando

$$\log \mu_l = x_l^\top \boldsymbol{\beta} + \tau \mathbb{I}(s_l \in Z), \tag{3.1}$$

em que $x_l = [x_{l1}, \ldots, x_{lp}]^\top$ é o vetor de covariáveis no local l, $\mathbb{I}(\cdot)$ é a função indicadora, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top$ é o vetor fixo de parâmetros desconhecidos, τ é o parâmetro de agrupamento, e μ_l é a média da variável resposta. Assim sendo,

$$\mu_l \equiv \begin{cases} \mu_{0l} = \exp\{x_l^\top \boldsymbol{\beta}\}, & \text{se} \quad s_l \notin Z; \\ \mu_{Zl} = \exp\{x_l^\top \boldsymbol{\beta} + \tau\}, & \text{caso contrário .} \end{cases}$$

Além disso, as hipóteses de interesse a serem contrastadas são $\mathcal{H}_0 : \tau = 0$ versus $\mathcal{H}_1 : \tau > 0$ (para algum *cluster* $Z \in \mathcal{Z}$). Nota-se que, sob a hipótese nula de $\tau = 0$, ou seja, de não existir *cluster* no mapa, o processo é modelado por $\mathcal{A}$-SCAN$(\mu_{0l}, \sigma, 0)$, $l = 1, \ldots, L$, e não depende do parâmetro τ. Por outro lado, sob a hipótese alternativa de que existe *cluster* no mapa $\mu_l \equiv \mu_{0l}$ se $s_l \notin Z$ e $\mu_l \equiv \mu_{Zl}$ se $s_l \in Z$. Em que, sob a hipótese $\mathcal{H}_1$ consideramos que $\mu_{Zl} > \mu_{0l}$. Consequentemente,

$$\exp\{\tau\} = \frac{\mu_{Zl}}{\mu_{0l}}.$$

Demonstração: Considere o caso em que $s_l \in Z$ logo,

$$\mu_{Zl} = \exp\{x_l^\top \boldsymbol{\beta} + \tau\} = \exp\{x_l^\top \boldsymbol{\beta}\} \times \exp\{\tau\}.$$

Substituindo $\mu_{0l} = \exp\{x_l^\top \boldsymbol{\beta}\}$, então:

$$\mu_{Zl} = \mu_{0l} \times \exp\{\tau\}.$$

Dessa forma, segue que:

$$\exp\{\tau\} = \frac{\mu_{Zl}}{\mu_{0l}}.$$

Portanto, o modelo $\mathcal{A}$-SCAN faz a comparação da média da variável resposta dos eventos que pertencem a zona Z versus a média da variável resposta dos eventos que não pertencem a zona Z por meio do parâmetro τ. Observa-se que, μ_{Zl} denota a média da variável resposta dos eventos que pertencem a zona Z e μ_{0l} é a média da variável resposta dos eventos que não pertencem a zona Z. Além disso, para efeitos comparativos o parâmetro τ indica na escala logarítmica a razão de chance ajustada por covariáveis para as observações Y_l que pertencem a zona Z em comparação com $Y_l's$ que não pertencem a zona Z.

3.2 Estatística de Teste para o Modelo de Regressão $\mathcal{A}$-SCAN

Considere a hipótese:

$$\mathcal{H}_0 : \tau = 0 \quad \text{vs} \quad \mathcal{H}_1 : \tau > 0 \text{ (para algum } \textit{cluster } Z \in \mathcal{Z}).$$

As hipóteses particulares de interesse são avaliar a significância do *cluster* detectado. A estatística de teste para esse tipo de teste é:

$$\Lambda = \max_{Z \in \mathcal{Z}} \hat{\Lambda}_Z, \quad \hat{\Lambda}_Z = \left\{ \log \ell_Z(\widehat{\boldsymbol{\theta}}_1, \hat{\tau}; \boldsymbol{y}) - \log \ell_0(\widehat{\boldsymbol{\theta}}_0, 0; \boldsymbol{y}) \right\}, \tag{3.2}$$

onde $\ell_Z(\cdot)$ representando a verossimilhança sob $\mathcal{H}_1$ para um conjunto particular de localização espacial Z e $\ell_0(\cdot)$ é a verossimilhança sob $\mathcal{H}_0$. Além disso, $\widehat{\boldsymbol{\theta}}_0 = [\widehat{\boldsymbol{\beta}}_0, \hat{\sigma}_0]^\top$ e $\hat{\tau}$ são os estimadores de máxima verossimilhança para os parâmetros do modelo de regressão sob as hipóteses nula e alternativa, respectivamente. Para este teste, sob $\mathcal{H}_1$, $\widehat{\boldsymbol{\theta}}_1 = [\widehat{\boldsymbol{\beta}}_1, \hat{\sigma}_1, \hat{\tau}]^\top$ com $\widehat{\boldsymbol{\beta}}_0 = \widehat{\boldsymbol{\beta}}_1$ e $\hat{\sigma}_0 = \hat{\sigma}_1$. Deve-se observar que os estimadores de máxima verossimilhança de $\boldsymbol{\beta}$ e σ assumem valores iguais para que o agrupamento deva ocorrer apenas pela variável resposta e não pela variação presente nas covariáveis.

3.3 Modelo Normal Inverso

A distribuição NI tem notável importância em diversas áreas de pesquisas. Logo, é fácil encontrar aplicações da distribuição normal inversa em áreas como análise de sobrevivência, confiabilidade, engenharia, estatística matemática. Além disso, essa distribuição é usada para descrever o primeiro tempo de passagem de uma partícula (movendo-se com velocidade constante) que está sujeita ao movimento browniano linear.

Seja Y uma variável aleatória com distribuição normal inversa (NI). Sob este modelo, a função densidade de probabilidade é dada por:

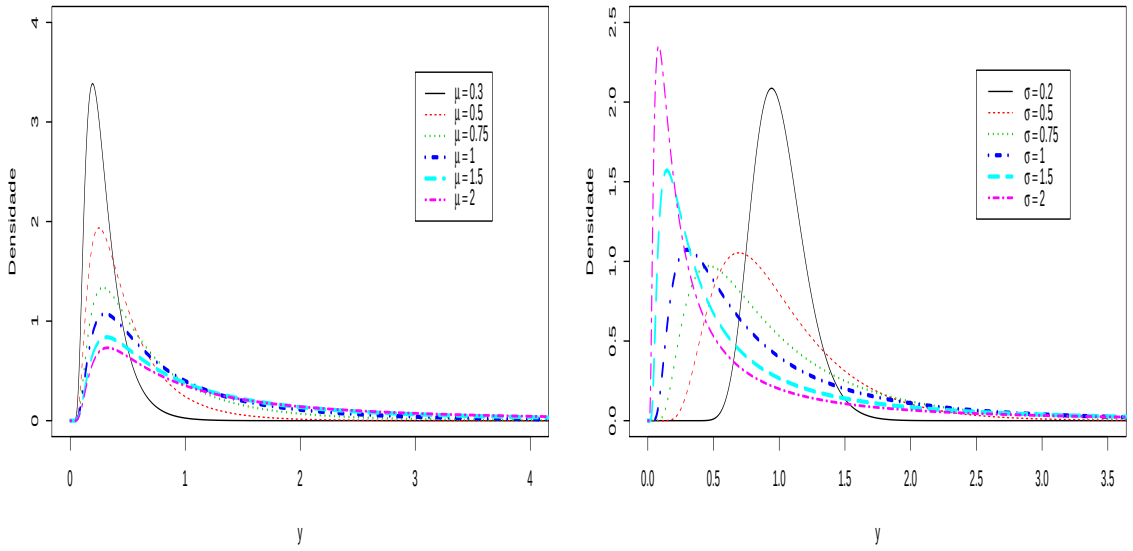
$$f(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left\{ -\frac{1}{2\mu^2\sigma^2 y} (y - \mu)^2 \right\},$$

para $y > 0$, $\mu > 0$ e $\sigma > 0$ e μ é a média e $\mu^3\sigma^2$ é a variância.

A Figura 3.1 apresenta as curvas de densidade do modelo NI considerando valores

para o parâmetro $\mu \in \{0,3; 0,5; 0,75; 1; 1,5; 2\}$ e σ fixo igual a 1. Portanto, nota-se que para valores menores de μ a curva de densidade do modelo apresenta uma forte assimetria. Também na Figura 3.1 ilustramos as curvas de densidade do modelo normal inverso considerando valores para o parâmetro $\sigma \in \{0,2; 0,5; 0,75; 1; 1,5; 2\}$ e μ fixo igual a 1. Logo, podemos observa que à medida que σ aumenta a curva de densidade do modelo vai ficando cada vez mais assimétrica e observa-se também por este gráfico que para valores pequenos de σ a densidade do modelo se aproxima da simetria em torno da média.

Figura 3.1: Densidade do modelo normal inverso para diferente valores de μ (a) e de σ (b).



(a) $\sigma = 1$ (b) $\mu = 1$

Fonte: Próprio autor

3.3.1 O Modelo NI-SCAN

Quando $\mathcal{A}(\mu, \sigma) \equiv \text{NI}(\mu, \sigma)$ e considerando o componente sistemático dado por (3.1), então, temos o modelo NI-SCAN$(\mu_l, \sigma, \tau), l = 1, \ldots, L$. Portanto, sob o modelo NI-SCAN(μ_l, σ, τ), quando $\tau > 0$, podemos escrever (3.2) como:

$$\hat{\Lambda}_Z^{NI} = \sum_{s_l \in Z} \hat{\Lambda}_l^{NI},$$

$$\hat{\Lambda}_l^{NI} = \frac{1}{\hat{\sigma}^2 \widehat{\mu}_{0l}} \left(\mathrm{e}^{-\hat{\tau}} - 1\right) \left\{1 - \frac{y_l}{2\widehat{\mu}_{0l}} \left(\mathrm{e}^{-\hat{\tau}} + 1\right)\right\}.$$

Logo, para $\tau > 0$ o *cluster* é estimado por $\hat{Z} = \arg\left(\max(\hat{\Lambda}_Z^{NI})\right)$; caso contrário, $\widehat{\Lambda}_Z^{NI} = 0$.

Demonstração: Apêndice A.

3.4 Modelo Gama

A distribuição GA proposta por (THOM, 1947), é um caso especial da distribuição de Pearson tipo III, onde o parâmetro de posição é zero. Este modelo é frequentemente usado em meteorologia e climatologia. Além disso, várias distribuições são casos particulares da distribuição gama por exemplo a exponencial, qui-quadrado e a Erlang.

A distribuição GA é um modelo probabilístico indexado por um parâmetro de locação μ e um parâmetro de escala σ, em que sua f.d.p pode ser escrita como:

$$f(y|\mu,\sigma) = \frac{y^{(1/\sigma^2-1)} \exp[-y/(\sigma^2\mu)]}{(\sigma^2\mu)^{(1/\sigma^2)}\Gamma(1/\sigma^2)},$$

para $y > 0$, $\mu > 0$ e $\sigma > 0$ será provado a seguir que μ é a média e $\mu^2\sigma^2$ é a variância dessa distribuição. Primeiramente seu k-ésimo momento é calculado da seguinte maneira:

$$\begin{aligned}
\mathbb{E}\left(Y^k\right) &= \int_0^\infty \frac{y^k y^{(1/\sigma^2-1)} \exp[-y/(\sigma^2\mu)]}{(\sigma^2\mu)^{(1/\sigma^2)}\Gamma(1/\sigma^2)} dy \\
&= \int_0^\infty \frac{y^{(1/\sigma^2+k-1)} \exp[-y/(\sigma^2\mu)]}{(\sigma^2\mu)^{(1/\sigma^2)}\Gamma(1/\sigma^2)} dy \\
&= \frac{1}{(\sigma^2\mu)^{(1/\sigma^2)}\Gamma(1/\sigma^2)} \times \int_0^\infty y^{(1/\sigma^2+k-1)} \exp[-y/(\sigma^2\mu)] dy.
\end{aligned}$$

Considere a seguinte substituição:

$$u = \frac{y}{(\sigma^2\mu)}, \;\; du = \frac{1}{(\sigma^2\mu)} dy, \;\; dy = (\sigma^2\mu)du, \;\; y = (\sigma^2\mu)u.$$

Logo,

$$\begin{aligned}\mathbb{E}\left(Y^k\right) =& \frac{1}{(\sigma^2\mu)^{(1/\sigma^2)}\Gamma(1/\sigma^2)} \times \int_0^\infty (u\sigma^2\mu)^{(1/\sigma^2+k-1)} e^{-u}(\sigma^2\mu)du \\ =& \frac{(\sigma^2\mu)^{(1/\sigma^2+k-1)}(\sigma^2\mu)}{(\sigma^2\mu)^{(1/\sigma^2)}\Gamma(1/\sigma^2)} \times \int_0^\infty u^{(1/\sigma^2+k-1)} e^{-u}\, du \\ =& \frac{(\sigma^2\mu)^{(1/\sigma^2)}(\sigma^2\mu)^k}{(\sigma^2\mu)^{(1/\sigma^2)}\Gamma(1/\sigma^2)} \times \int_0^\infty u^{(1/\sigma^2+k-1)} e^{-u}\, du \\ =& \frac{(\sigma^2\mu)^k}{\Gamma(1/\sigma^2)} \times \int_0^\infty u^{(1/\sigma^2+k-1)} e^{-u}\, du\end{aligned}$$

Portanto,

$$\mathbb{E}\left(Y^k\right) = \frac{(\sigma^2\mu)^k}{\Gamma(1/\sigma^2)} \times \Gamma(1/\sigma^2 + k).$$

Consequentemente,

$$\begin{aligned}\mathbb{E}\left(Y\right) =& \frac{(\sigma^2\mu)}{\Gamma(1/\sigma^2)} \times \Gamma(1/\sigma^2 + 1) \\ =& \frac{(\sigma^2\mu)}{\Gamma(1/\sigma^2)} \times \frac{1}{\sigma^2} \times \Gamma(1/\sigma^2) \\ =& \mu.\end{aligned}$$

Para demonstrar que $Var(Y) = \sigma^2\mu^2$, primeiro devemos calcula a $\mathbb{E}(Y^2)$. Por meio da função de momentos o segundo momento da distribuição gama é dado por:

$$\mathbb{E}\left(Y^2\right) = \mu^2 + \sigma^2\mu^2.$$

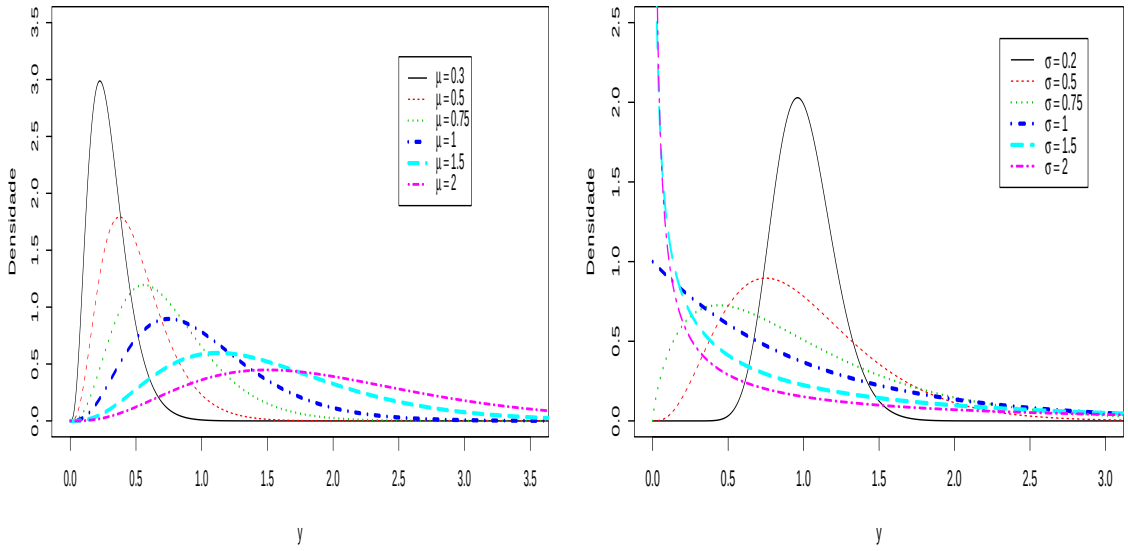
Portanto,

$$\begin{aligned}\mathrm{Var}(Y) =& \mathbb{E}\left(Y^2\right) - \mathbb{E}^2\left(Y\right) \\ =& \mu^2 + \sigma^2\mu^2 - \mu^2 \\ =& \sigma^2\mu^2.\end{aligned}$$

A Figura 3.2(a) apresenta as curvas de densidade do modelo gama considerando valores para o parâmetro de locação $\mu \in \{0,3; 0,5; 0,75; 1; 1,5; 2\}$ e σ fixo igual a $0,5$. Logo, podemos observa que há uma mudança na variância do modelo, ou seja, à medida que μ aumenta a variância também aumenta. Na Figura 3.2(b) ilustramos as

curvas de densidade do modelo gama considerando valores para o parâmetro de escala $\sigma \in \{0,2; 0,5; 0,75; 1; 1,5; 2\}$ e μ fixo igual a 1. Assim sendo, nota-se que à medida que o parâmetro de escala σ decresce a densidade do modelo se aproxima da simetria em torno da média.

Figura 3.2: Densidade do modelo gama para diferentes valores de μ(a) e de σ(b).



(a) $\sigma = 0,5$ (b) $\mu = 1$

Fonte: Próprio autor

3.4.1 O Modelo GA-SCAN

Aqui $\mathcal{A}(\mu, \sigma) \equiv \text{GA}(\mu, \sigma)$ e se considerarmos o componente sistemático dado por (3.1), portanto, temos o modelo GA-SCAN$(\mu_l, \sigma, \tau), l = 1, \ldots, L$. Considerando que estamos sob o modelo GA-SCAN(μ_l, σ, τ), quando $\tau > 0$, então podemos escrever (3.2) como:

$$\widehat{\Lambda}_Z^{GA} = \sum_{s_l \in Z} \hat{\Lambda}_l^{GA},$$

$$\hat{\Lambda}_l^{GA} = \frac{1}{\hat{\sigma}^2}\left[\frac{y_l(1-e^{-\hat{\tau}})}{\hat{\mu}_{0l}} - \hat{\tau}\right].$$

Portanto, quando rejeitamos $\mathcal{H}_0$ o *cluster* é estimado por $\hat{Z} = \arg\left(\max(\hat{\Lambda}_Z^{GA})\right)$; caso contrário, $\widehat{\Lambda}_Z^{GA} = 0$.

Demonstração: Apêndice B.

3.5 Modelo Weibull

A distribuição WE tem notável importância e portanto podemos encontrar aplicações dela em diversas áreas de conhecimento, tais como: na análise de sobrevivência, física, biologia e engenharia de confiabilidade. Além disso, a distribuição WE é uma opção atraente para modelagem de sobrevivência totalmente paramétrica, uma vez que, unicamente, ela possui o tempo de falha acelerado e a propriedade de riscos proporcionais. De modo geral, as aplicações visam a determinação do tempo de vida médio e da taxa de falhas em função do tempo da população observada. Um detalhe histórico é que a distribuição recebeu tal denominação devido a Waloddi Weibull que em 1951 publicou um artigo descrevendo a distribuição em detalhes e propondo diversas aplicações.

A distribuição WE é uma distribuição de probabilidade contínua indexada por um parâmetro de forma ($\sigma > 0$) e um parâmetro de escala ($\beta > 0$) e sua função densidade de probabilidade pode ser escrita da seguinte maneira:

$$f(y|\sigma,\beta) = \frac{\sigma}{\beta^{\sigma}} y^{(\sigma-1)} \exp\left\{-\left(\frac{y}{\beta}\right)^{\sigma}\right\}, \quad y,\sigma,\beta \in (0,\infty),$$

e sua função de distribuição acumulada é escrita da seguinte forma:

$$F(Y) = \begin{cases} 0, & \text{se } y \ 0; \\ 1 - \exp\left[-\left(\frac{y}{\beta}\right)^{\sigma}\right], & \text{se } y \ \geq \ 0. \end{cases}$$

Seu k-ésimo momento é calculado do seguinte modo:

$$\mathbb{E}\left(Y^k\right) = \int_0^{\infty} y^k \frac{\sigma}{\beta^{\sigma}} y^{\sigma-1} e^{-(y/\beta)^{\sigma}} dy,$$

seja,

$$u = \left(\frac{y}{\beta}\right)^{\sigma}, \qquad du = \frac{1}{\beta^{\sigma}} \sigma y^{\sigma-1} dy, \qquad dy = \frac{\beta^{\sigma}}{\sigma y^{\sigma-1}} du.$$

Dessa forma,

$$\begin{aligned} \mathbb{E}\left(Y^k\right) &= \int_0^{\infty} \beta^k u^{k/\sigma} \frac{\sigma}{\beta^{\sigma}} y^{\sigma-1} e^{-u} \frac{\beta^{\sigma}}{\sigma y^{\sigma-1}} du \\ &= \int_0^{\infty} \beta^k u^{k/\sigma} e^{-u} du \\ &= \beta^k \int_0^{\infty} u^{k/\sigma+1-1} e^{-u} du. \end{aligned}$$

Logo,

$$\mathbb{E}\left(Y^k\right) = \beta^k \Gamma\left(\frac{k}{\sigma} + 1\right).$$

Consequentemente,

$$\mathbb{E}\left(Y\right) = \beta \Gamma\left(\frac{1}{\sigma} + 1\right). \tag{3.3}$$

Calculando a variância da distribuição Weibull, vamos primeiro calcular a $\mathbb{E}\left(Y^2\right)$. Por meio da função de momentos tem-se que:

$$\mathbb{E}\left(Y^2\right) = \beta^2 \Gamma\left(\frac{2}{\sigma} + 1\right).$$

Portanto,

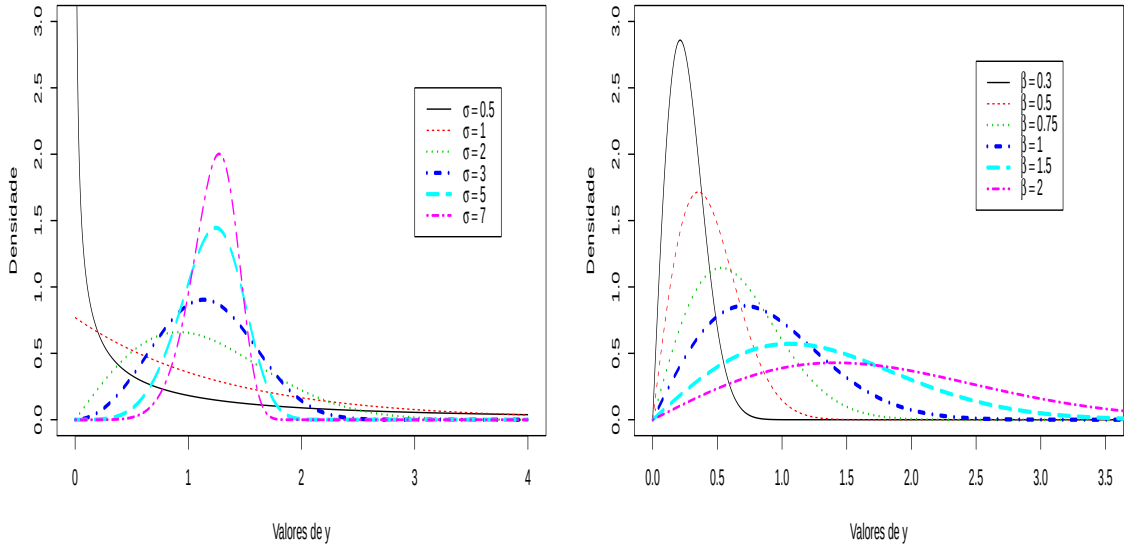
$$\begin{aligned} \text{Var}(Y) =& \mathbb{E}\left(Y^2\right) - \mathbb{E}^2\left(Y\right) \\ =& \beta^2 \Gamma\left(\frac{2}{\sigma} + 1\right) - \left(\beta \Gamma\left(\frac{1}{\sigma} + 1\right)\right)^2. \end{aligned}$$

Logo,

$$\text{Var}(Y) = \beta^2 \left\{\Gamma\left(1 + \frac{2}{\sigma}\right) - \Gamma^2\left(\frac{1}{\sigma} + 1\right)\right\}. \tag{3.4}$$

A Figura 3.3-(a) apresenta as curvas de densidade do modelo Weibull considerando valores para o parâmetro de forma $\sigma \in \{0,5; 1; 2; 3; 5; 7\}$ e β fixo igual a $1,3$. Logo, nota-se que variações no valor do parâmetro de forma σ alteram bastante o comportamento do modelo. Alem disso, algumas variações no parâmetro σ faz a equação do modelo se reduzir a outros modelos de probabilidade. Na Figura 3.3-(b) ilustramos as curvas de densidade do modelo Weibull considerando valores para o parâmetro de escala $\beta \in \{0,3; 0,5; 0,75; 1; 1,5; 2\}$ e α fixo igual a 2. Assim sendo, nota-se que há uma mudança na média e variância do modelo Weibull, ou seja, à medida que β aumenta os valores da média e variância crescem.

Figura 3.3: Densidade do modelo Weibull para diferente valores de σ e de β.



(a) $\beta = 1,3$ (b) $\sigma = 2$

Fonte: Próprio autor

Usando uma nova parametrização para a distribuição Weibull em que $\beta = \frac{\mu}{\Gamma(1+\frac{1}{\sigma})}$ desse modo a função de densidade de probabilidade (f.d.p) da distribuição Weibull é dada por:

$$f(y|\mu,\sigma) = \frac{\sigma}{\left(\frac{\mu}{\Gamma(1+\frac{1}{\sigma})}\right)^{\sigma}} y^{(\sigma-1)} \exp\left\{-\left(\frac{y\Gamma(1+\frac{1}{\sigma})}{\mu}\right)^{\sigma}\right\}, \quad y,\mu,\sigma \in (0,\infty).$$

Sob essa nova parametrização substituindo $\beta = \frac{\mu}{\Gamma(1+\frac{1}{\sigma})}$ na expressão (3.3) obtemos $\mathbb{E}(Y) = \mu$.

Além disso, substituindo $\beta = \frac{\mu}{\Gamma(1+\frac{1}{\sigma})}$ na expressão (3.4) então concluímos que,

$$\text{Var}[Y] = \mu^2 \left\{ \frac{\Gamma(1+\frac{2}{\sigma})}{\left[\Gamma^2(1+\frac{1}{\sigma})\right]} - 1 \right\}.$$

3.5.1 O Modelo WE-SCAN

Agora tem-se que $\mathcal{A}(\mu,\sigma) \equiv \text{WE}(\mu,\sigma)$ e se considerarmos o componente sistemático dado por (3.1), logo, temos o modelo WE-SCAN$(\mu_l,\sigma,\tau), l = 1,\ldots,L$. Portanto,

sob o modelo WE-SCAN(μ_l, σ, τ), quando $\tau > 0$, podemos escrever (3.2) como:

$$\widehat{\Lambda}_Z^{WE} = \sum_{s_l \in Z} \hat{\Lambda}_l^{WE},$$

$$\hat{\Lambda}_l^{WE} = \left[\frac{y_l \Gamma(1 + 1/\hat{\sigma})}{\widehat{\mu}_{0l}}\right]^{\hat{\sigma}} (1 - \mathrm{e}^{-\hat{\tau}\hat{\sigma}}) - \hat{\sigma}\hat{\tau}.$$

Assim sendo, quando rejeitamos $\mathcal{H}_0$ o *cluster* é estimado por $\hat{Z} = \arg\left(\max(\hat{\Lambda}_Z^{WE})\right)$; caso contrário, $\widehat{\Lambda}_Z^{WE} = 0$.

Demonstração: Apêndice C.

3.6 Modelo Birnbaum-Saunders

A distribuição BS proposta em Birnbaum e Saunders (1969), no artigo intitulado *a new family of life distributions*, tem recebido considerável atenção nos últimos anos, devido aos seus argumentos teóricos associados aos processos de danos cumulativos, suas propriedades e sua relação com a distribuição normal. Comumente chamada de distribuição de vida por fadiga, tem sido amplamente utilizada para análise de dados de sobrevivência e também em aplicações nas áreas de engenharia, confiabilidade, medicina, finanças, agricultura, indústria e em diversas outras áreas. Por exemplo, a distribuição BS é indicada para modelar o tempo até a ocorrência de falha em processos de fadiga quando estamos interessados nos percentis mais baixos ou mais altos da distribuição, ou seja, nas caldas da distribuição, pois nesse caso a distribuição BS tem um ajuste satisfatório. Por outro lado, os modelos probabilístico GA, WE, NI e lognormal são indicados para modelar o tempo até a ocorrência de falha em processos de fadiga quando estamos interessados na região central da distribuição de vida pois esses modelos se ajustam bem nesta situação. Especificamente, presume-se que a quantidade de dano cumulativo que permite que a distribuição BS seja gerada segue uma distribuição normal. Este modelo corresponde a uma distribuição unimodal, inclinada positivamente, de dois parâmetros e com suporte não negativo.

A fadiga ocorre quando um material é exposto a situações de estresse e tensão, assim espera-se naturalmente que este material deverá sofrer danos ou rachaduras em sua estrutura e, a esses danos denominamos de fadiga. Segundo Santos-Neto (2010), motivados por problemas nos aviões comerciais novos e falhas de material, Birnbaum

e Saunders (1969) derivaram uma nova família de distribuições de vida que modela o tempo de vida de materiais e equipamentos submetidos a cargas dinâmicas. Aqui apresentaremos as suposições propostas por Birnbaum e Saunders (1969) sobre o processo de fadiga, veja Santos-Neto (2010). Essas suposições são as seguintes:

(i) Um material é sujeito a um padrão cíclico de tensão e força;

(ii) A sequência de tensão imposta ao material é a constante de ciclo para ciclo;

(iii) A falha por fadiga do material ocorre devido ao desenvolvimento e ao crescimento de uma rachadura dominante dentro do material. Portanto, a falha ocorre quando o tamanho da rachadura dominante excede certo nível de resistência ω;

(iv) A extensão incremental da rachadura X_i resultante da aplicação da i-ésima oscilação de carga é uma variável aleatória com uma distribuição que só dependerá da rachadura atual causada pela tensão neste ciclo;

(v) A extensão da rachadura durante o $(j+1)$-ésimo ciclo é

$$Y_{j+1} = X_{jm+1} + \cdots + X_{jm+m}, \qquad j = 0, 1, 2, \ldots,$$

em que, X_{jm+1} é a extensão da rachadura após a i-ésima oscilação de carga do $(j+1)$-ésimo ciclo;

(vi) As extensões das rachaduras em diferentes ciclos são diferentes;

(vii) A extensão total da rachadura, Y_j devido ao j-ésimo ciclo é uma variável aleatória com distribuição de média μ e variância σ^2, $\forall\ j = 1, 2, \ldots$;

Assim sendo, a extensão total da rachadura após n ciclos será dada pela variável aleatória,

$$W_n = \sum_{j=1}^{n} Y_j,$$

com função de distribuição acumulada (f.d.a) escrita do seguinte modo:

$$H_n(w) = \mathbb{P}(W_n \leq \omega), \qquad \forall\ n = 1, 2, \ldots .$$

Seja N a variável aleatória que representa o número de ciclos necessários até que ocorra uma falha. Deste modo, a função de distribuição acumulada de N é dada por:

$$\mathbb{P}(N \leq n) = \mathbb{P}(\sum_{j=1}^{n} Y_j > \omega) = \mathbb{P}(W_n > \omega) = 1 - H_n(\omega).$$

Imagine que os $Y_j's$ sejam variáveis aleatórias independentes e identicamente distribuídas (i.i.d). Logo, segue do teorema central do limite que,

$$\begin{aligned}
\mathbb{P}(N \leq n) =& \mathbb{P}\left(\sum_{j=1}^{n} \frac{Y_j - \mu}{\sigma\sqrt{n}} > \frac{\omega - n\mu}{\sigma\sqrt{n}}\right) = 1 - \mathbb{P}\left(\sum_{j=1}^{n} \frac{Y_j - \mu}{\sigma\sqrt{n}} \leq \frac{\omega - n\mu}{\sigma\sqrt{n}}\right) \\
=& 1 - \mathbb{P}\left(\sum_{j=1}^{n} \frac{Y_j - \mu}{\sigma\sqrt{n}} \leq \frac{\omega}{\sigma\sqrt{n}} - \frac{\mu\sqrt{n}}{\sigma}\right) \cong 1 - \Phi\left(\frac{\omega}{\sigma\sqrt{n}} - \frac{\mu\sqrt{n}}{\sigma}\right) \\
\cong& \Phi\left[-\left(\frac{\omega}{\sigma\sqrt{n}} - \frac{\mu\sqrt{n}}{\sigma}\right)\right] \cong \Phi\left(\frac{\mu\sqrt{n}}{\sigma} - \frac{\omega}{\sigma\sqrt{n}}\right), \qquad (3.5)
\end{aligned}$$

em que $\Phi(\cdot)$ denota a função de distribuição normal padrão.

Utilizando a equação (3.5), Birnbaum e Saunders (1969) definiram uma nova distribuição de vida. Segundo Birnbaum e Saunders (1969), quando substituirmos n por uma variável real não negativa t, consequentemente a variável aleatória T será a extensão contínua da variável discreta N. Logo, T representa o tempo total até a ocorrência da falha. Considere $\alpha = \sigma/\sqrt{\mu\omega}$ e $\beta = \omega/\mu$, então a (f.d.a) de T é escrita do seguinte modo:

$$F(t|\alpha,\beta) = \mathbb{P}(T \leq t) = \Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right)\right],\ t > 0, \alpha > 0, \beta > 0, \qquad (3.6)$$

em que $\Phi(\cdot)$ é a função de distribuição da normal padrão. Usaremos a notação $T \sim BS(\alpha,\beta)$ para indicar que T é uma variável aleatória seguindo uma distribuição BS com parâmetros α e β que são os parâmetros de forma e escala respectivamente. Além disso, β é a mediana da distribuição, isto é, $F(\beta|\alpha,\beta) = \Phi(0) = 0,5$.

A f.d.p da distribuição de T obtida através de (3.6) pode ser escrita da seguinte maneira:

$$f(t|\alpha,\beta) = \frac{\exp\{\alpha^{-2}\}}{2\alpha\sqrt{2\pi\beta}}\, t^{-3/2}\,[t+\beta]\exp\left\{-\frac{1}{2\alpha^2}\left[\frac{t}{\beta} + \frac{\beta}{t}\right]\right\}, \quad t > 0, \alpha > 0, \beta > 0. \qquad (3.7)$$

A média é a variância associadas com (3.7) são dadas respectivamente por:

$$\mathbb{E}(T) = \beta\left(1 + \frac{1}{2}\alpha^2\right) \quad \text{e} \quad \text{Var}(T) = (\alpha\beta)^2\left(1 + \frac{5}{4}\alpha^2\right).$$

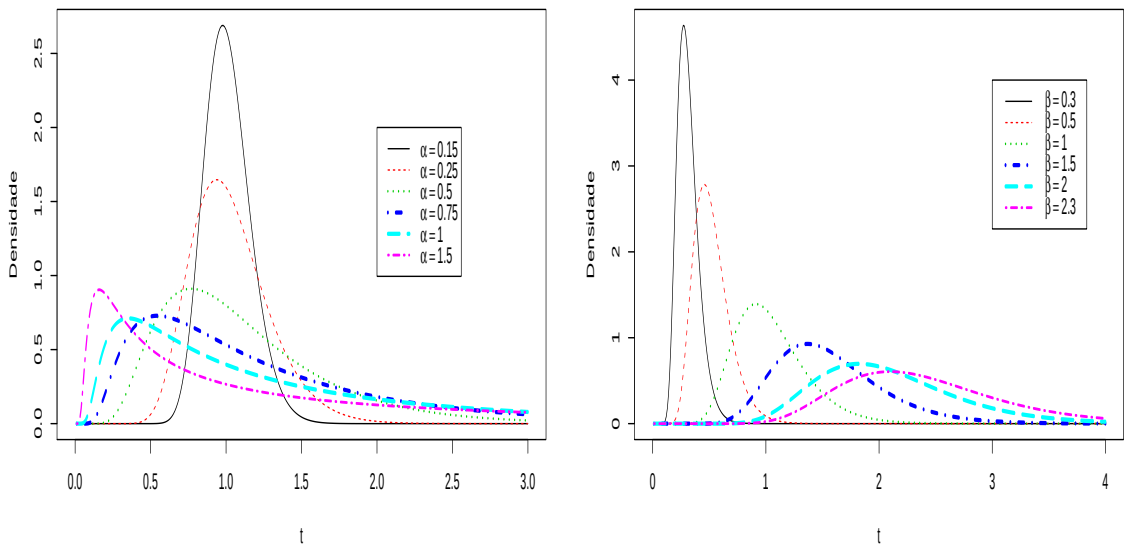
A distribuição $BS(\alpha, \beta)$ está relacionada com a distribuição normal, está relação é a seguinte:

$$T = \beta\left[\frac{\alpha Z}{2} + \sqrt{\left(\frac{\alpha Z}{2}\right)^2 + 1}\right]^2,$$

em que $Z \sim N(0,1), \alpha > 0$ e $\beta > 0$.

A Figura 3.4(a) apresenta as curvas de densidade do modelo BS considerando diferentes valores para o parâmetro de forma α e β fixo igual a 1. Portanto, de acordo com a Figura 3.4 observa-se que à medida que o valor de α aumenta altera a forma da curva da densidade, ou seja, a forma da curva da densidade se torna cada vez mais assimétrica. Já na Figura 3.4(b) variamos os valores do parâmetro β e consideramos α fixo. Podemos observar que há uma mudança na média e variância do modelo BS, ou seja, à medida que β aumenta os valores da média e variância também aumentam.

Figura 3.4: Densidade Birnbaum-Saunders para diferentes valores de α e de β.



(a) $\beta = 1$ (b) $\alpha = 0,3$

Fonte: Próprio autor

Santos-Neto et al. (2012) propuseram várias parametrizações para a distribuição BS usando diferentes argumentos. Um deles indexa a distribuição BS por sua média e

precisão, que denominamos de BSR. Leiva et al. (2015), Leiva (2016), Santos-Neto et al. (2016) e Leão et al. (2017) mostraram que a distribuição BSR é útil em configurações para as quais a parametrização original é limitada.

A f.d.p da distribuição de BSR é dada por:

$$f(y|\mu,\sigma) = \frac{\exp(\sigma/2)\sqrt{\sigma+1}}{4\sqrt{\pi\mu}\, y^{3/2}}\left[y + \frac{\sigma\mu}{\sigma+1}\right]\exp\left(-\frac{\sigma}{4}\left[\frac{\{\sigma+1\}y}{\sigma\mu} + \frac{\sigma\mu}{\{\sigma+1\}y}\right]\right),$$

em que $y > 0$, $\sigma > 0$ e $\mu > 0$ são parâmetros de forma (precisão) e escala (média), respectivamente.

A média e variância da distribuição BSR são dadas respectivamente por:

$$\mathbb{E}(Y) = \mu, \quad \text{e} \quad \text{Var}(Y) = \frac{\text{g}(\mu)}{\text{h}(\sigma)},$$

em que,

$$\text{g}(\mu) = 2\mu^2 \quad e \quad \text{h}(\sigma) = \frac{1}{\dfrac{1}{\sigma+2+\dfrac{1}{\sigma}} + \dfrac{5}{2(\sigma^2+\sigma+1)}}.$$

3.6.1 O Modelo BSR-SCAN

Se $\mathcal{A}(\mu,\sigma) \equiv \text{BSR}(\mu,\sigma)$ e considerando o componente sistemático dado por (3.1), então temos o modelo BSR-SCAN$(\mu_l,\sigma,\tau), l = 1,\ldots,L$. Sob o modelo BSR-SCAN$(\mu_l,\sigma,\tau)$, quando $\tau > 0$, podemos escrever (3.2) como:

$$\widehat{\Lambda}_Z^{BSR} = \sum_{s_l \in Z} \hat{\Lambda}_l^{BSR},$$

$$\hat{\Lambda}_l^{BSR} = -\frac{\hat{\tau}}{2} + \log\left[\frac{\hat{\sigma}y_l + y_l + \hat{\sigma}\hat{\mu}_{Zl}}{\hat{\sigma}y_l + y_l + \hat{\sigma}\hat{\mu}_{0l}}\right] + \frac{y_l\{\hat{\sigma}+1\}}{4\hat{\mu}_{0l}}(1-e^{-\hat{\tau}}) + \frac{\hat{\sigma}^2\hat{\mu}_{0l}}{4\{\hat{\sigma}+1\}y_l}(1-e^{\hat{\tau}}).$$

Desta maneira, quando $\tau > 0$ o *cluster* é estimado por $\hat{Z} = \arg\left(\max(\hat{\Lambda}_Z^{BSR})\right)$; caso contrário, $\widehat{\Lambda}_Z^{BSR} = 0$.

Demonstração: Apêndice D.

3.7 Modelo Beta-Prime

A distribuição BP (KEEPING, 1962; MCDONALD, 1984) é uma distribuição alternativa de tempo de falha, como as distribuições BS, GA, NI e WE. Assim, podemos

encontrar aplicações da distribuição BP na área de análise de sobrevivência e também em aplicações nas áreas de engenharia, confiabilidade, medicina e finanças. Uma variável aleatória T segue a distribuição BP com os parâmetros de forma $\alpha > 0$ e $\beta > 0$, denotados pela BP(α, β), se a distribuição de T admitir a seguinte densidade em relação à medida de Lebesgue:

$$f(t|\alpha,\beta) = \frac{t^{\alpha-1}(1+t)^{-(\alpha+\beta)}}{B(\alpha,\beta)}, \quad t > 0, \tag{3.8}$$

sua função de distribuição acumulada (f.d.a) é dada por:

$$F(t|\alpha,\beta) = I_{\frac{t}{1+t}}(\alpha,\beta), \quad t > 0,$$

em que $I_y(\alpha,\beta) = B_y(\alpha,\beta)/B(\alpha,\beta)$ é a razão incompleta da função beta, $B_y(\alpha,\beta) = \int_0^y u^{\alpha-1}(1-u)^{\beta-1}\mathrm{d}u$ é a função incompleta, $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ é a função beta e $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}\exp(-u)\mathrm{d}u$ é a função gama. A média e a variância associadas com (3.8) são dadas por:

$$\mathbb{E}(T) = \frac{\alpha}{\beta-1}, \quad \beta > 1, \quad \text{e} \quad \mathrm{Var}(T) = \frac{\alpha(\alpha+\beta-1)}{(\beta-2)(\beta-1)^2}, \quad \beta > 2,$$

respectivamente.

Demonstração:

$$\begin{aligned}\mathbb{E}(T) &= \int_0^\infty t \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{t^{\alpha-1}}{(1+t)^{\alpha+\beta}}dt \\ &= \int_0^\infty \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{t^{\alpha}}{(1+t)^{\alpha+\beta}}dt \end{aligned} \tag{3.9}$$

É possível demonstrar que $T = \dfrac{Y}{1-Y}$ tem distribuição BP de acordo com a equação (3.8), por exemplo ver Bourguignon, Santos-Neto e Castro (2018). Sendo assim, considere a seguinte substituição:

$$t = \frac{y}{1-y} \qquad \Rightarrow \qquad y = \frac{t}{1+t},$$

observa-se que,

$$t+1 = \frac{1}{1-y} \qquad \text{e} \qquad dy = \frac{dt}{(1+t)^2} \qquad \Rightarrow \qquad dt = (1+t)^2 dy\cdot$$

Aplicando a substituição acima em (3.9). Isto é,

$$
\begin{aligned}
\mathbb{E}(T) &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\left(\frac{y}{1-y}\right)^{\alpha}}{\left(\frac{1}{1-y}\right)^{\alpha+\beta}} \times \frac{dy}{(1+t)^{-2}} \\
&= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\left(\frac{y}{1-y}\right)^{\alpha}}{\left(\frac{1}{1-y}\right)^{\alpha+\beta}} \times \frac{dy}{\left(\frac{1}{1-y}\right)^{-2}} \\
&= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\left(\frac{y}{1-y}\right)^{\alpha}}{\left(\frac{1}{1-y}\right)^{\alpha}\left(\frac{1}{1-y}\right)^{\beta-2}}\, dy \\
&= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \left(\frac{\frac{y}{1-y}}{\frac{1}{1-y}}\right)^{\alpha} \times \frac{1}{\frac{1}{(1-y)^{\beta-2}}}\, dy \\
&= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times y^{\alpha} \times (1-y)^{\beta-2}\, dy \\
&= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times y^{(\alpha+1)-1} \times (1-y)^{(\beta-1)-1}\, dy \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 y^{(\alpha+1)-1} \times (1-y)^{(\beta-1)-1}\, dy \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(\alpha+1)\Gamma(\beta-1)}{\Gamma(\alpha+1+\beta-1)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(\alpha+1)\Gamma(\beta-1)}{\Gamma(\alpha+\beta)} \\
&= \frac{\alpha\Gamma(\alpha)\Gamma(\beta-1)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \frac{\alpha\Gamma(\beta-1)}{\Gamma(\beta-1+1)} \\
&= \frac{\alpha\Gamma(\beta-1)}{(\beta-1)\Gamma(\beta-1)} \\
&= \frac{\alpha}{(\beta-1)}, \qquad \beta > 1.
\end{aligned}
$$

Agora será calculado a $\mathbb{E}(T^2)$.

Demonstração:

$$\begin{aligned}\mathbb{E}(T^2) &= \int_0^\infty t^2 \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{t^{\alpha-1}}{(1+t)^{\alpha+\beta}} dt \\ &= \int_0^\infty \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{t^{\alpha+1}}{(1+t)^{\alpha+\beta}} dt. \end{aligned} \tag{3.10}$$

Considere a seguinte substituição:

$$t = \frac{y}{1-y} \qquad \Rightarrow \qquad y = \frac{t}{1+t},$$

observa-se que,

$$t+1 = \frac{1}{1-y} \qquad \text{e} \qquad dy = \frac{dt}{(1+t)^2} \qquad \Rightarrow \qquad dt = (1+t)^2 dy.$$

Aplicando a substituição acima em (3.10). Ou seja,

$$\begin{aligned}\mathbb{E}(T^2) &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\left(\frac{y}{1-y}\right)^{\alpha+1}}{\left(\frac{1}{1-y}\right)^{\alpha+\beta}} \times \frac{dy}{(1+t)^{-2}} \\ &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\left(\frac{y}{1-y}\right)^{\alpha+1}}{\left(\frac{1}{1-y}\right)^{\alpha+\beta}} \times \frac{dy}{\left(\frac{1}{1-y}\right)^{-2}} \\ &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\left(\frac{y}{1-y}\right)^{\alpha+1}}{\left(\frac{1}{1-y}\right)^{\alpha+1}\left(\frac{1}{1-y}\right)^{\beta-3}} \, dy \\ &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \left(\frac{\frac{y}{1-y}}{\frac{1}{1-y}}\right)^{\alpha+1} \times \frac{1}{\frac{1}{(1-y)^{\beta-3}}} \, dy \\ &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times y^{\alpha+1} \times (1-y)^{\beta-3} \, dy \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 y^{\alpha+1+1-1} \times (1-y)^{\beta-3+1-1} \, dy \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 y^{(\alpha+2)-1} \times (1-y)^{(\beta-2)-1} \, dy \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(\alpha+2)\Gamma(\beta-2)}{\Gamma(\alpha+2+\beta-2)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(\alpha+2)\Gamma(\beta-2)}{\Gamma(\alpha+\beta)} \\
&= \frac{\Gamma(\alpha+2)\Gamma(\beta-2)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \frac{(\alpha+1)\Gamma(\alpha+1)\Gamma(\beta-2)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \frac{(\alpha+1)\alpha\Gamma(\alpha)\Gamma(\beta-2)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \frac{\alpha(\alpha+1)\Gamma(\beta-2)}{\Gamma(\beta)} \\
&= \frac{\alpha(\alpha+1)\Gamma(\beta-2)}{\Gamma(\beta-2+2)}.
\end{aligned}
$$

Faça $\eta = \beta - 2$, então:

$$
\begin{aligned}
\mathbb{E}(T^2) &= \frac{\alpha(\alpha+1)\Gamma(\eta)}{\Gamma(\eta+2)} \\
&= \frac{\alpha(\alpha+1)\Gamma(\eta)}{(\eta+1)\eta\Gamma(\eta)} \\
&= \frac{\alpha(\alpha+1)}{(\eta+1)\eta} \\
&= \frac{\alpha(\alpha+1)}{(\beta-1)(\beta-2)}, \qquad \beta > 2.
\end{aligned}
$$

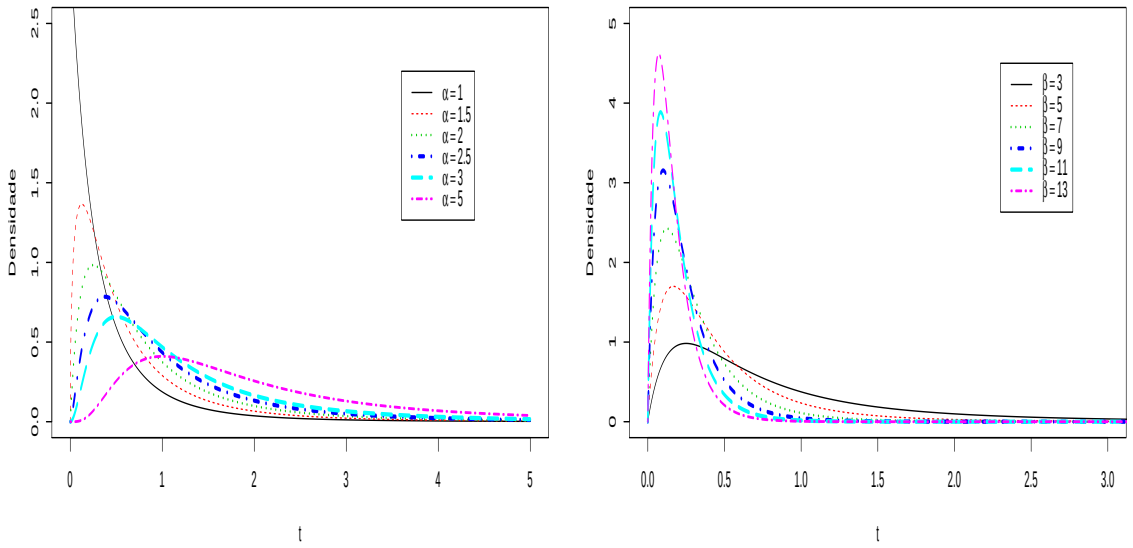
Portanto,

$$
\begin{aligned}
\mathrm{Var}(T) &= \mathbb{E}\left(T^2\right) - \mathbb{E}^2\left(T\right) \\
&= \frac{\alpha(\alpha+1)}{(\beta-1)(\beta-2)} - \left(\frac{\alpha}{\beta-1}\right)^2 \\
&= \frac{\alpha(\alpha+\beta-1)}{(\beta-2)(\beta-1)^2}, \quad \beta > 2.
\end{aligned}
$$

A Figura 3.5(a) apresenta as curvas de densidade do modelo Beta-Prime considerando valores para o parâmetro de forma $\alpha \in \{1; 1,5; 2; 2,5; 3; 5\}$ e β fixo igual a 3. Assim, nota-se que à medida que α aumenta há um deslocamento para a direita da densidade do modelo, isto é, um aumento na média e variância do modelo. Na Figura 3.5(b) ilustramos as curvas de densidade do modelo Beta-Prime considerando valores para o parâmetro de escala $\beta \in \{3; 5; 7; 9; 11; 13\}$ e α fixo igual a 2. Logo, podemos observa que à medida que β aumenta a média e variância do modelo diminui e nota-se

também que, neste caso, a densidade do modelo vai ficando cada vez mais concentrada em torno do zero e com caudas mais leve.

Figura 3.5: Densidade do modelo Beta-Prime para diferentes valores de α e β.



(a) $\beta = 3$ (b) $\alpha = 2$

Fonte: Próprio autor

Bourguignon, Santos-Neto e Castro (2018) define uma estrutura de regressão para respostas distribuídas pela BP usando uma nova parametrização que difere de (3.8). Seja, em (3.8), $\alpha = \mu(\sigma + 1)$ e $\beta = \sigma + 2$, isto é, $\alpha = \mu(1 + \sigma)$ e $\beta = 2 + \sigma$. Sob essa nova parametrização, se $T \sim \text{BP}(\alpha, \beta)$ então $\mathbb{E}(T) = \mu$ e $\text{Var}(T) = \mu(1 + \mu)/\sigma$; nesta parametrização, usaremos a notação $T \sim \text{BP}(\mu, \sigma)$ para indicar que T é uma variável aleatória seguindo uma distribuição BP com média μ e parâmetro de precisão σ. Usando essa nova parametrização, a densidade da BP em (3.8) pode ser escrito como:

$$f(t|\mu, \sigma) = \frac{t^{\mu(\sigma+1)-1}(1+t)^{-[\mu(\sigma+1)+\sigma+2]}}{B(\mu(\sigma+1), (\sigma+2))}, \tag{3.11}$$

em que $\mu > 0$ e $\sigma > 0$, uma vez que $\alpha > 0$ e $\beta > 0$.

3.7.1 O Modelo BP-SCAN

Quando $\mathcal{A}(\mu, \sigma) \equiv \text{BP}(\mu, \sigma)$ e considerando o componente sistemático dado por (3.1), então temos o modelo BP-SCAN$(\mu_l, \sigma, \tau), l = 1, \ldots, L$. Sob o modelo

BP-SCAN(μ_l, σ, τ), quando $\tau > 0$, podemos escrever (3.2) como:

$$\widehat{\Lambda}_Z^{BP} = \sum_{s_l \in Z} \hat{\Lambda}_l^{BP},$$

$$\hat{\Lambda}_l^{BP} = \hat{\mu}_{0l}(\hat{\sigma}+1)(e^{\hat{\tau}}-1)\log\left(\frac{y_l}{1+y_l}\right) + \log\left(\frac{B(\hat{\mu}_{0l}(1+\hat{\sigma}),(2+\hat{\sigma}))}{B(\hat{\mu}_{zl}(1+\hat{\sigma}),(2+\hat{\sigma}))}\right).$$

Portanto, quando rejeitamos $\mathcal{H}_0$ o *cluster* é estimado por $\hat{Z} = \arg\left(\max(\hat{\Lambda}_Z^{BP})\right)$; caso contrário, $\widehat{\Lambda}_Z^{BP} = 0$.

Demonstração: Apêndice E.

3.8 O Teste do p-valor Via *Bootstrap* para a Estatística Espacial $\mathcal{A}$-SCAN

Como a distribuição exata da estatística de teste Λ é desconhecida, ou seja, é analiticamente intratável sob $\mathcal{H}_0$, sendo assim, podemos usar o p-valor do *Bootstrap* para avaliar a significância estatística do *cluster* mais provável. Esta técnica é aplicada nesta situação pois os parâmetros da distribuição da estatística $\mathcal{A}$-SCAN são desconhecidos sob a hipótese nula, então uma maneira de contornar este tipo de problema é a técnica de *Bootstrap*, deste modo, podemos estimar uma distribuição obtida a partir dos dados, denominada como função de distribuição empírica. O Método *Bootstrap* consiste em um método de reamostragem que foi proposto por Efron (1979). O algoritmo utilizado é basicamente o algoritmo descrito em Lima et al. (2016).

O algoritmo do p-valor do *Bootstrap* para avaliar a significância estatística de Λ consiste nos seguintes passos:

Passo 1: Com base na amostra e nas covariáveis calcular $\widehat{\boldsymbol{\theta}}_0$ e $\hat{\tau}$. Derive o valor observado de Λ e denote por $\widehat{\Lambda}$;

Passo 2: Gerar amostras de *Bootstrap* $\boldsymbol{y}_b^* = (y_{1,b}^*, \ldots, y_{L,b}^*)$ de $\mathcal{A}$-SCAN$(\mu_{0l}(\hat{\beta}_0), \hat{\sigma}, 0), l = 1, 2, \ldots, L$;

Passo 3: Do passo 2 calcular $\widehat{\boldsymbol{\theta}}_0^*$ e derive $\widehat{\Lambda}_b^*$;

Passo 4: Repita os passos 2 e 3 para $b = 1, \ldots, B-1$ e calcule o p-valor para Λ por meio de $\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\widehat{\Lambda} \geq \widehat{\Lambda}_b^*)$.

Capítulo 4

Estudo de Simulação

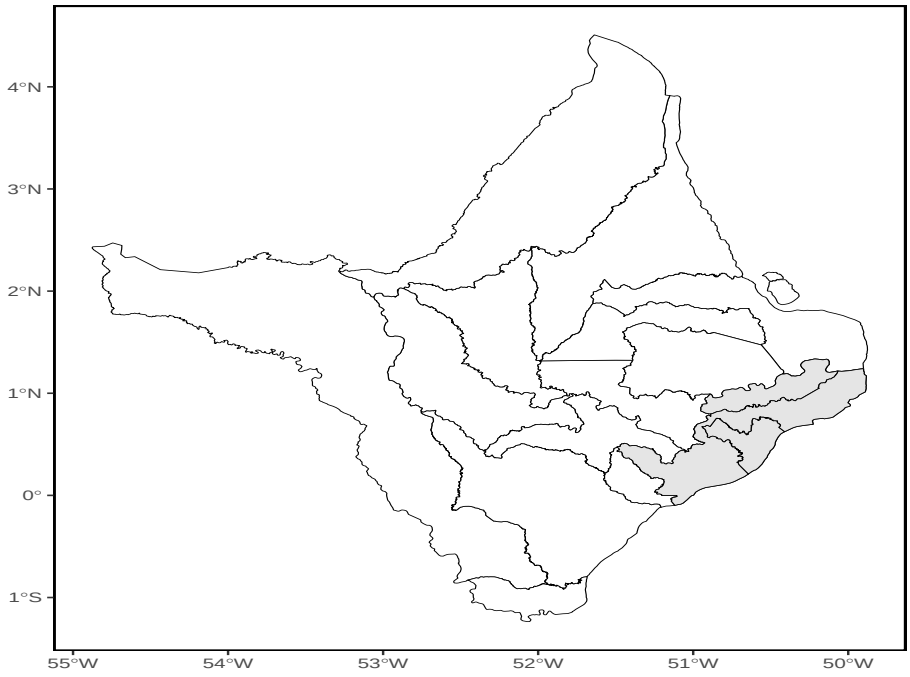
Neste Capítulo, são apresentados os resultados das simulações realizadas para avaliar a performance das novas metodologias usando a estatística *scan* circular. Consideramos como região de estudo o mapa do estado do Amapá no Brasil. A escolha do mapa se deve ao fato de ser um estado com apenas 16 municípios, proporcionando uma redução significativa no tempo das simulações. Pois, quanto maior a quantidade de regiões de um mapa, maior o tempo computacional necessário para executar o algoritmo da estatística scan espacial. Acreditamos que a escolha desse mapa não represente perda importante na qualidade de nossas simulações.

Na Figura 4.1, apresentamos o mapa do estado do Amapá com o *cluster* artificial considerado nas simulações. A seleção do *cluster* foi feita após a realização de várias simulações considerando diferentes *clusters*. O poder de detecção do *cluster* depende da geração dos dados e dos diferentes valores dos parâmetros que compõem o modelo. Na comparação entre as estatísticas *scan* espacial baseadas nos modelos de regressão BSR, BP, GA, NI e WE, serão verificados os comportamentos dos gráficos do poder do teste, do valor preditivo positivo e da sensibilidade. Desta forma, esperamos definir qual distribuição apresentou melhor desempenho na identificação do *cluster* artificial.

Nos modelos NI-SCAN e GA-SCAN o parâmetro σ é um parâmetro de dispersão, diferentemente dos demais modelos, em que este parâmetro é classificado como de precisão. Para garantir a mesma interpretação, consideramos para os modelo NI e GA o inverso de σ. Além disso, é importante destacar que o modelo NI tem uma tendência de

apresentar uma maior variabilidade, pois sua variância além de ser função de σ também é função de um termo cúbico da média, enquanto que nos demais modelos o termo da média é função quadrática. Desta forma, espera-se que o modelo NI-SCAN apresente um pior desempenho durante as simulações.

Figura 4.1: *Cluster* artificial com 3 áreas (Itaubal, Cutias e Macapá).



Fonte: Próprio autor

Para garantir que durante as simulações os cenários fossem os mesmos para todos os modelos, foram gerados os mesmos valores para as variáveis preditoras, possibilitando assim que para todos os modelos as médias geradas fossem iguais. Na Tabela 4.1, mostramos os intervalos para as médias e variâncias para todos os modelos considerando $\tau = 10$. Além disso, calculamos a variância média e a mediana das variâncias. Como esperado, nota-se que a variabilidade dos modelos reduz com o aumento do valor de σ. Também pode-se notar que o modelo NI-SCAN apresenta a maior variância média em todos os cenários. Nota-se que para o modelo BSR-SCAN com $\sigma = 20$, 50% das variâncias são menores ou iguais a 105,71. De uma maneira geral, no cenário com $\sigma = 0,5$, os modelos BSR-SCAN e BP-SCAN apresentaram as menores médias e medianas das variâncias. Já para $\sigma = 2,20$, esses valores foram menores nos modelos GA-SCAN e WE-SCAN.

Tabela 4.1: Intervalos de variação das médias e das variâncias para $\tau = 10$.

Modelos	σ	μ_i	$\text{Var}[Y_i]$	$\overline{\text{Var}[Y_i]}$	$\text{Med}(\text{Var}[Y_i])$
BSR-SCAN	0,5	(2,28, 686,81)	(13,83, 1257880,27)	94386,51	2762,55
	2	(2,28, 686,81)	(5,19, 471705,10)	35394,94	1035,96
	20	(2,28, 686,81)	(0,53, 48133,17)	3611,73	105,71
BP-SCAN	0,5	(2,28, 686,81)	(14,93, 944783,82)	70967,83	2134,62
	2	(2,28, 686,81)	(3,73, 236195,95)	17741,96	533,65
	20	(2,28, 686,81)	(0,37, 23619,60)	1774,20	53,37
GA-SCAN	0,5	(2,28, 686,81)	(20,75, 1886820,40)	141579,77	4143,83
	2	(2,28, 686,81)	(1,3, 117926,3)	8848,74	258,99
	20	(2,28, 686,81)	(0,01, 1179,26)	88,49	2,59
NI-SCAN	0,5	(2,28, 686,81)	(47,27, 1295883000)	85895415,37	143194,21
	2	(2,28, 686,81)	(2,95, 80992698,04)	5368463,46	8949,64
	20	(2,28, 686,81)	(0,03, 809926,98)	53684,64	89,50
WE-SCAN	0,5	(2,28, 686,81)	(25,94, 2358525,50)	176974,71	5179,79
	2	(2,28, 686,81)	(1,42, 128888,49)	9671,30	283,06
	20	(2,28, 686,81)	(0,02, 1811,85)	135,95	3,98

Fonte: Próprio autor

Inicialmente, sob a hipótese nula, consideramos 1000 réplicas de Monte Carlo (devido ao custo computacional alto) para obter valores críticos empíricos do teste com o nível de significância $\alpha = 0,05$ para os modelos BSR-SCAN$(\mu_{0,l}, \sigma, 0)$, BP-SCAN$(\mu_{0,l}, \sigma, 0)$, GA-SCAN$(\mu_{0,l}, \sigma, 0)$, NI-SCAN$(\mu_{0,l}, \sigma, 0)$ e WE-SCAN$(\mu_{0,l}, \sigma, 0)$, $l = 1, \ldots, 16$ com

$$\mu_{0,l} = \frac{\exp\{0,734 + 4x_l\}}{1 + \exp\{0,734 + 4x_l\}}, \quad \sigma = 0,5,2,20 \quad \text{e} \quad x_l \sim U(0,1).$$

Na Tabela 4.2 apresentamos os valores críticos, para o nível de significância de 5%, considerando as distribuições empíricas das estatísticas de teste. Nas Figuras 4.2-4.5 mostramos as formas das distribuições empíricas das estatísticas de teste para os diferentes valores de σ. Notamos que as distribuições geralmente apresentam uma assimetria positiva.

Tabela 4.2: Valores críticos empíricos obtidos a partir das distribuições empíricas sob a hipótese nula para diferentes estatísticas e valores de σ.

Modelos	σ	Valores Críticos Empíricos
BSR-SCAN	0,5	15,216
	2	15,017
	20	3,4741
BP-SCAN	0,5	13,605
	2	13,482
	20	3,653
GA-SCAN	0,5	8,189
	2	14,768
	20	3,556
NI-SCAN	0,5	10447,240
	2	15,011
	20	3,151
WE-SCAN	0,5	9,238
	2	18,767
	20	2,136

Fonte: Próprio autor.

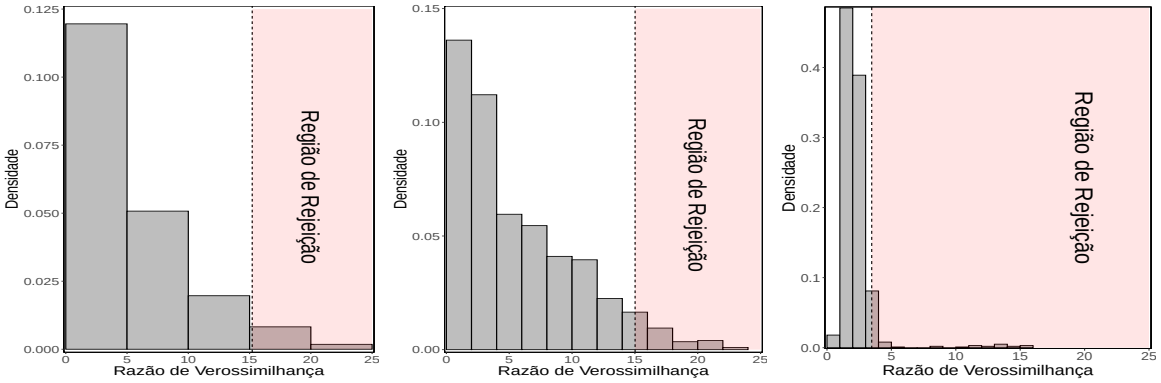
Para o cálculo do poder do teste, consideramos os valores críticos obtidos sob a hipótese nula ($\tau = 0$) apresentados na Tabela 4.2. Isto garante que todas as estatísticas tenham o mesmo tamanho do teste. Sob a hipótese alternativa consideramos, os valores de $\tau = \log(i), i = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$. Os comportamentos do poder do teste (a), Sensibilidade (b) e Valor Predito Positivo (c) para os modelos BSR-SCAN, GA-SCAN, BP-SCAN, NI-SCAN e WE-SCAN são apresentados nas Figuras 4.6, 4.7, 4.8, 4.9 e 4.10, respectivamente.

Importante destacar que para o valor de $\sigma = 0,5$ no modelo NI-SCAN, o valor crítico empírico para a estatística de teste Λ^{NI} foi muito grande o que impossibilitou a criação do histograma de modo informativo. Desta forma, também decidimos omitir no texto os histogramas das distribuições empíricas da estatística Λ^{NI} sob a hipótese nula para $\sigma = \{2, 20\}$, com o objetivo de mantermos o mesmo padrão de apresentação dos gráficos neste capítulo. Além disso, de acordo com a Tabela 4.2 nota-se que para o modelo NI-SCAN quando o valor do parâmetro σ cresce os valores críticos empíricos para a estatística Λ^{NI} decresce.

4.1 Análise dos Resultados

Na Figura 4.2, apresentamos a distribuição empírica da estatística Λ^{BSR} sob a hipótese nula para $\sigma = \{0,5, 2, 20\}$. De acordo com a Tabela 4.2 e Figura 4.2, pode-se observar que à medida que o valor do parâmetro σ aumenta os valores críticos empíricos para a estatística Λ^{BSR} decresce. Por exemplo, para $\sigma = 0,5$ o valor crítico empírico foi 15,216, por outro lado, para $\sigma = 20$ o valor crítico empírico foi 3,4741. É importante observar que a estatística de teste Λ^{BSR} depende de σ e mudança no parâmetro σ leva a essa variação na estatística de teste Λ^{BSR}.

Figura 4.2: Distribuição empírica da estatística Λ^{BSR} sob a hipótese nula para $\sigma = \{0,5, 2, 20\}$.

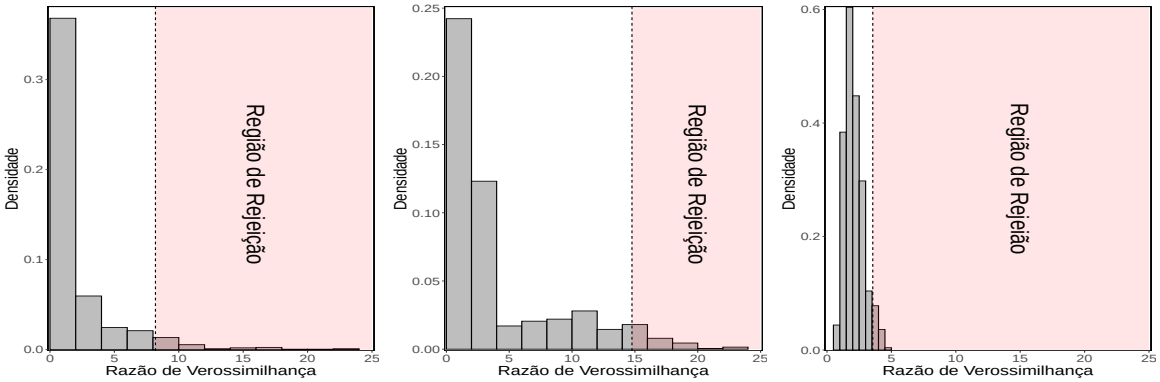


(a) $\sigma = 0,5$ (b) $\sigma = 2$ (c) $\sigma = 20$

Fonte: Próprio autor

Na Figura 4.3 apresentamos a distribuição empírica da estatística Λ^{GA} sob a hipótese nula para $\sigma = \{0,5, 2, 20\}$. Nota-se que, de acordo com a Tabela 4.2 e Figura 4.3, que o menor valor crítico empírico para Λ^{GA} foi 3,556 para $\sigma = 20$ e maior valor crítico empírico para Λ^{GA} foi 14,768 para $\sigma = 2$, quando diminuímos o valor de σ para 0,5 o valor crítico empírico obtido foi de 8,189. Adicionalmente, a estatística de teste Λ^{GA} depende de σ e mudança no parâmetro σ causa essa variação na estatística de teste Λ^{GA}.

Figura 4.3: Distribuição empírica da estatística Λ^{GA} sob a hipótese nula para $\sigma = \{0,5, 2, 20\}$.

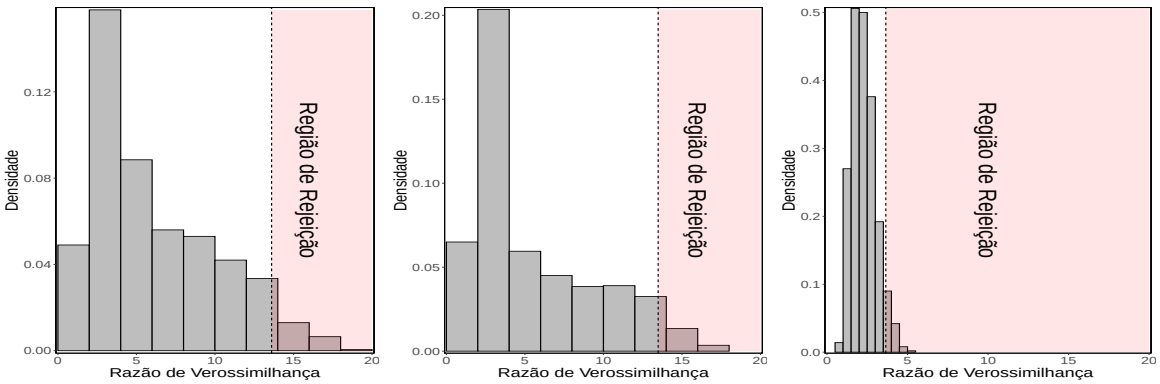


(a) $\sigma = 0,5$ (b) $\sigma = 2$ (c) $\sigma = 20$

Fonte: Próprio autor

Na Figura 4.4, ilustramos a distribuição empírica da estatística Λ^{BP} sob a hipótese nula para $\sigma = \{0,5, 2, 20\}$. Diante do exposto na Figura 4.4 e na Tabela 4.2 podemos observar que à medida que o valor do parâmetro σ aumenta os valores críticos empíricos para a estatística Λ^{BP} diminuíram. Por exemplo, para $\sigma = 0,5$ o valor crítico empírico foi 13,605, por outro lado, para $\sigma = 20$ o valor crítico empírico foi 3,653. Esse comportamento da estatística de teste Λ^{BP} pode está associado ao fato de Λ^{BP} depende do parâmetro σ e também por causa que à medida que o valor de σ aumenta a variância do modelo Beta-Prime (Equação 3.11) diminui.

Figura 4.4: Distribuição empírica da estatística Λ^{BP} sob a hipótese nula para $\sigma = \{0,5, 2, 20\}$.

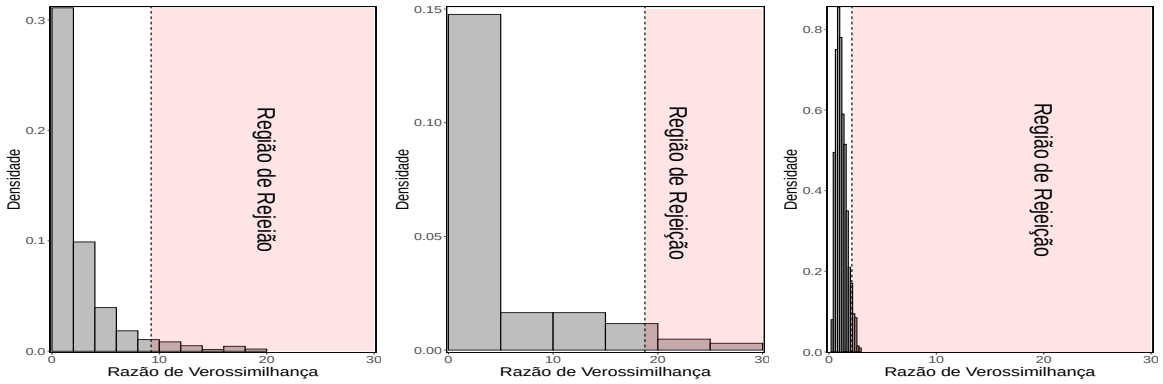


(a) $\sigma = 0,5$ (b) $\sigma = 2$ (c) $\sigma = 20$

Fonte: Próprio autor

Na Figura 4.5 apresentamos a distribuição empírica da estatística Λ^{WE} sob a hipótese nula para $\sigma = \{0,5, 2, 20\}$. Percebe-se que, de acordo com a Tabela 4.2 e Figura 4.5, que o menor valor crítico empírico para Λ^{WE} foi 2,136 para $\sigma = 20$ e maior valor crítico empírico para Λ^{WE} foi 18,767 para $\sigma = 2$, quando diminuímos o valor de σ para 0,5 o valor crítico empírico obtido foi de 9,238. Nota-se que, a estatística de teste Λ^{WE} depende de σ e mudança no parâmetro σ causa essa variação na estatística de teste Λ^{WE}.

Figura 4.5: Distribuição empírica da estatística Λ^{WE} sob a hipótese nula para $\sigma = \{0,5, 2, 20\}$.

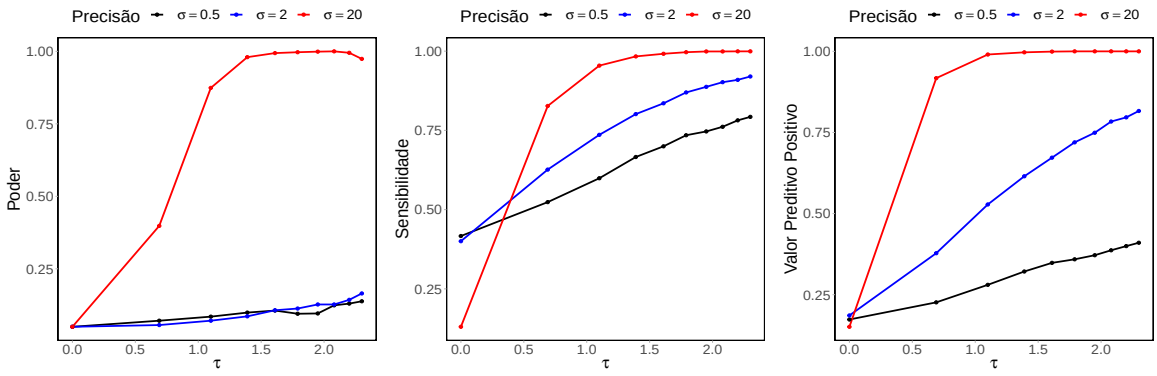


(a) $\sigma = 0,5$ (b) $\sigma = 2$ (c) $\sigma = 20$

Fonte: Próprio autor

Podemos notar na Figura 4.6 que à medida que o parâmetro τ aumenta o poder do teste cresce isto é mais claro para $\sigma = 20$, nota-se também que com o aumento do valor do parâmetro τ a sensibilidade e valor predito positivo também aumentam para o modelo BSR-SCAN. Observa-se que, o poder do teste é muito baixo para valores de $\sigma = \{0,5, 2\}$, porém para $\sigma = 20$ o método apresenta um bom poder de teste e uma boa precisão para detectar o local correto do *cluster*, pois , neste caso, os valores das medidas de sensibilidade e valor predito positivo estão acima de 0,75. Portanto, o método BSR-SCAN tem uma boa precisão para detectar o *cluster* verdadeiro para $\sigma = 20$. Além disso, percebe-se que o poder do teste, SS e VPP crescem com o aumento de σ. Portanto, fica claro que o poder de detecção do *cluster* depende dos diferentes valores dos parâmetros que compõem o modelo.

Figura 4.6: Gráficos da Função poder (a), Sensibilidade (b) e Valor Predito Positivo (c) para o modelo BSR-SCAN.

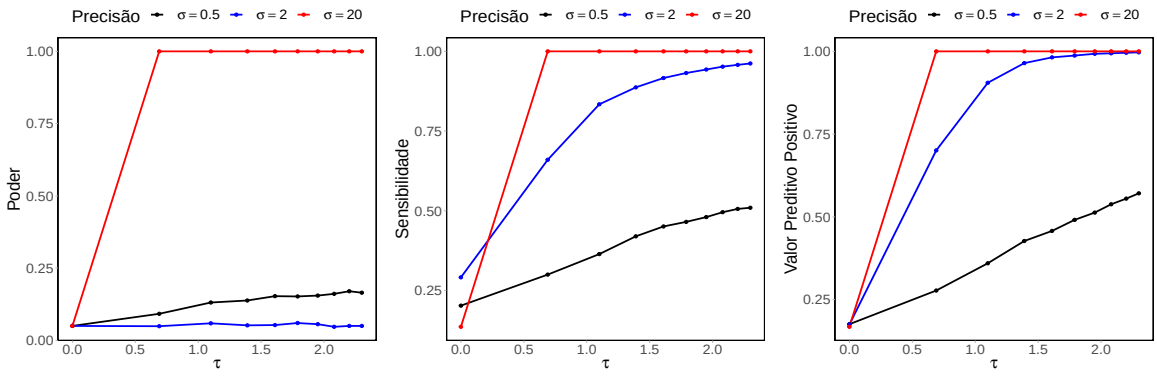


(a) (b) (c)

Fonte: Próprio autor

Na Figura 4.7 para o modelo GA-SCAN, notamos que quando o valor do parâmetro τ aumenta os valores das medidas de sensibilidade e valor predito positivo também aumentam. Percebe-se que, o poder do teste é muito pequeno para $\sigma = \{0, 5, 2\}$. No entanto, para $\sigma = 20$ o método apresenta um alto poder de teste e uma alta precisão para detectar o local correto do *cluster*, pois os valores das medidas de sensibilidade e valor predito positivo crescem juntos para 1. Além do mais, quando aumentamos o valor do parâmetro σ a sensibilidade e valor predito positivo crescem. Assim sendo, ficou claro o efeito da variação dos parâmetros σ e τ no desempenho do método GA-SCAN para detectar o *cluster* verdadeiro.

Figura 4.7: Gráficos da Função poder (a), Sensibilidade (b) e Valor Predito Positivo (c) para o modelo GA-SCAN.

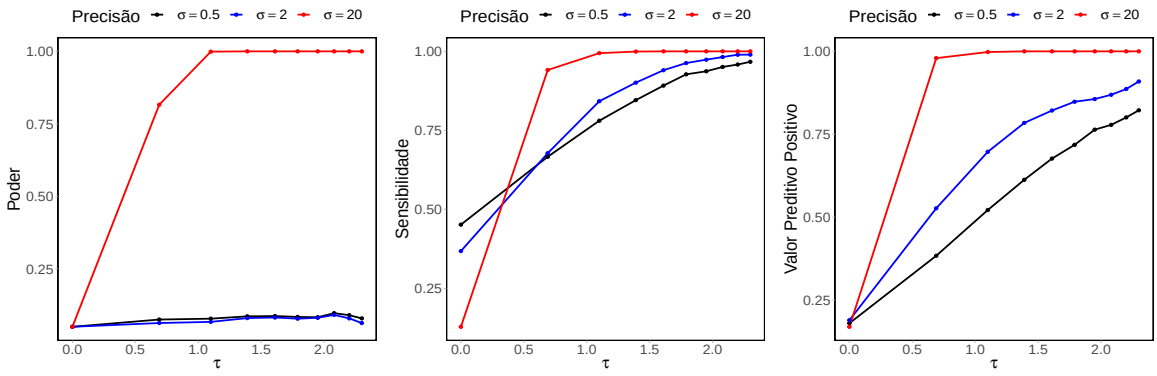


(a) (b) (c)

Fonte: Próprio autor

Diante do exposto na Figura 4.8 para o modelo BP-SCAN, observa-se que quando o valor do parâmetro τ cresce a sensibilidade e valor predito positivo aumentam. Adicionalmente, nota-se que o poder do teste é muito pequeno para $\sigma = \{0, 5, 2\}$. No entanto, para $\sigma = 20$ o método apresenta um bom poder de teste e uma alta precisão para detectar o local correto do *cluster*, pois os valores das medidas de sensibilidade e valor predito positivo crescem juntos para valores mais próximos de 1. Além do mais, quando aumentamos o valor do parâmetro σ a sensibilidade e valor predito positivo crescem e o poder do teste foi maior para $\sigma = 20$ isto indicar que com o aumento de σ o poder do teste aumenta. Diante disso, podemos visualizar que o poder de detecção do *cluster* depende dos diferentes valores dos parâmetros σ e τ.

Figura 4.8: Gráficos da Função poder (a), Sensibilidade (b) e Valor Predito Positivo (c) para o modelo BP-SCAN.

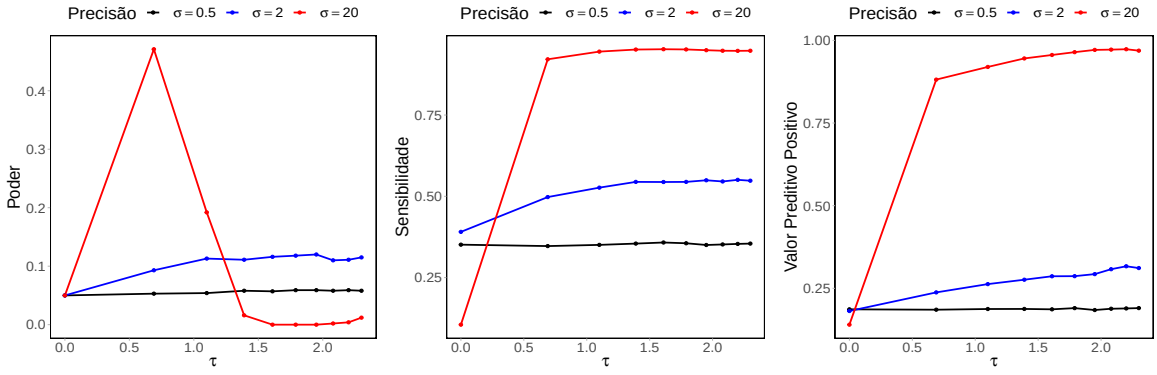


(a) (b) (c)

Fonte: Próprio autor

De acordo com a Figura 4.9 para o modelo NI-SCAN, observa-se que o poder do teste é muito pequeno para $\sigma = \{0, 5, 2, 20\}$. Para $\sigma = 20$ o poder do teste apresentou um comportamento inadequado, pois o esperado na literatura e que com o aumento dos parâmetros σ e τ o poder do teste aumente. Portanto, o modelo NI-SCAN apresentou um desempenho muito baixo para detectar o *cluster* artificial para todos os valores de σ e τ considerados nesta simulação.

Figura 4.9: Gráficos da Função poder (a), Sensibilidade (b) e Valor Predito Positivo (c) para o modelo NI-SCAN.



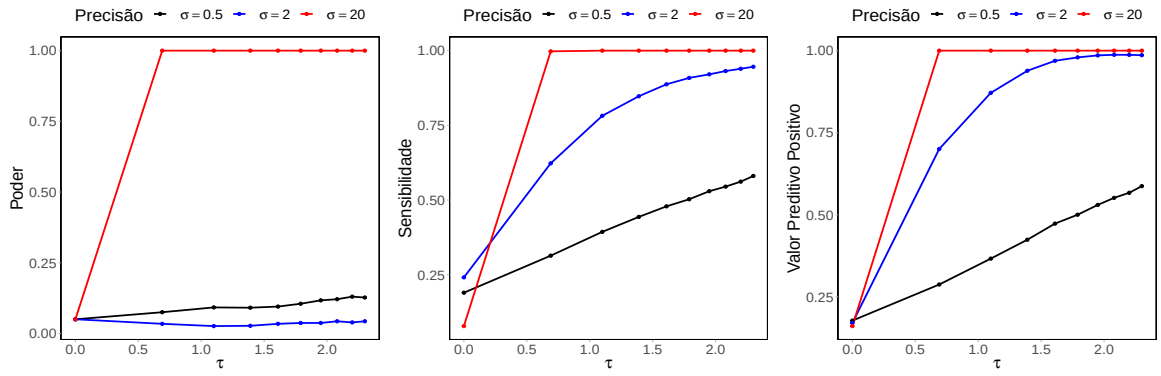
(a) (b) (c)

Fonte: Próprio autor

Na Figura 4.10 para o modelo WE-SCAN, notamos que quando o valor do parâmetro τ cresce o valor das medidas de sensibilidade e valor predito positivo também

crescem. De maneira análoga aos demais modelos utilizados nas simulações observa-se que, o poder do teste é muito pequeno para $\sigma = \{0, 5, 2\}$. Mas, para $\sigma = 20$ o método apresenta um alto poder de teste e uma alta precisão para detectar o local correto do *cluster* pois, os valores das medidas de sensibilidade e valor predito positivo crescem juntos para 1. Além disso, quando aumentamos o valor do parâmetro σ a sensibilidade e valor predito positivo crescem. Logo, fica evidente por esses resultados que o poder de detecção do *cluster* depende dos diferentes valores dos parâmetros que compõem o modelo.

Figura 4.10: Gráficos da Função poder (a), Sensibilidade (b) e Valor Predito Positivo (c) para o modelo WE-SCAN.



(a) (b) (c)

Fonte: Próprio autor

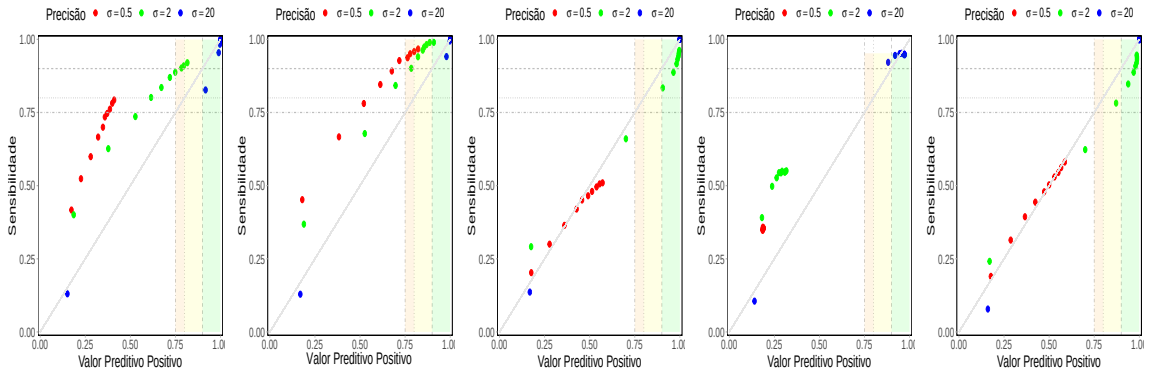
Analisando a Figura 4.11 percebe-se que as Figuras 4.11 (a) e (b) apresentaram comportamentos semelhantes a única diferença é que para o método BP-SCAN os valores de SS e VPP são maiores e estão mais próximos da reta $y = x$, isto indica que o método BP-SCAN tem uma melhor precisão para detectar o local correto do *cluster* quando ele existe, do que o BSR-SCAN. Nota-se também para esses dois modelos que para valores de $\sigma = \{0, 5, 2\}$ a SS é maior que VPP, ou seja, os métodos BP-SCAN e BSR-SCAN tendem a superestimar o *cluster* verdadeiro neste cenário. Por outro lado, para $\sigma = 20$ VPP é maior que SS, ou seja, os métodos BP-SCAN e BSR-SCAN tendem a subestimar o *cluster* verdadeiro neste cenário, mas os valores de SS e VPP são próximos de 1 indicando uma boa precisão para detectar o local correto do *cluster*.

As Figuras 4.11 (c) e (e) também apresentaram comportamentos semelhantes, para $\sigma = 0.5$ os valores de SS e VPP estão próximos da reta y=x mas não mostra uma

boa precisão para detectar o local correto do *cluster*, pois os valores dessas medidas são baixos. Para $\sigma = 2$ VPP é maior que SS, indicando que os métodos GA-SCAN e WE-SCAN tende a subestimar o *cluster* neste cenário. Para $\sigma = 20$ os valores de SS e VPP estão próximos da reta $y = x$, além disso, os valores dessas medidas estão mais próximo de 1, isto mostra que os métodos GA-SCAN e WE-SCAN tem alta precisão para detectar o local correto do *cluster* neste cenário.

Na Figura 4.11 (d) notamos para o método NI-SCAN que ele apresentou valores de SS e VPP menores que os demais modelos considerando todos os cenários, ou seja, para $\sigma = \{0, 5, 2, 20\}$ e nestes cenários SS é maior que VPP, indicando que o método tende a superestimar o *cluster* verdadeiro, além disso, os valores de SS e VPP foram pequenos para $\sigma = \{0, 5, 2\}$ isto indica que o método não mostra uma boa precisão para detectar o local correto do *cluster* neste cenário. Para $\sigma = 20$ se a hipótese nula for rejeitada o método tem uma boa precisão para detectar o local correto do *cluster* quando ele existe, pois os valores de SS e VPP estão acima de 0,75.

Figura 4.11: Gráficos da sensibilidade contra o valor preditivo positivo para os modelos BSR-SCAN (a), BP-SCAN (b), GA-SCAN (c), NI-SCAN (d) e WE-SCAN (e).



(a) BSR-SCAN (b) BP-SCAN (c) GA-SCAN (d) NI-SCAN (e) WE-SCAN

Fonte: Próprio autor

Para efeitos comparativos, de modo geral o método NI-SCAN apresentou um pior desempenho em relação aos demais métodos utilizados na simulação, pois os valores das medidas de poder do teste, sensibilidade e valor preditivo positivo foram menores em comparação aos valores dessas medidas para os demais modelos. O BSR-SCAN apresentou um desempenho inferior ao BP-SCAN. Notamos que para valores de $\sigma = \{0, 5, 2\}$ todos os modelos apresentaram poder de teste baixo, além disto, para

este cenário os métodos BSR-SCAN, BP-SCAN e NI-SCAN tendem a superestimar o *cluster* verdadeiro já os métodos GA-SCAN e WE-SCAN tendem a subestimar o *cluster* verdadeiro neste cenário. Diante do exposto na Figura 4.11, percebe-se que o método BP-SCAN os valores de SS e VPP são maiores em comparação aos demais modelos considerando todos os cenários, ou seja, para $\sigma = \{0, 5, 2, 20\}$. Isto indica que o método BP-SCAN tem uma melhor precisão para detectar o local correto do *cluster* quando ele existe do que os demais modelos. Logo, o modelo BP-SCAN foi escolhido como o melhor modelo para detecção de *cluster* é será utilizado na parte da aplicação. Além disso, o BP-SCAN apresentou um bom poder do teste e uma alta precisão para detectar o *cluster* quando ele existe considerando o cenário com $\sigma = 20$.

Capítulo 5

Aplicação

Neste Capítulo, é apresentada a análise realizada nos dados provenientes do estado do Maranhão e relativos a pacientes com tuberculose. O objetivo é identificar regiões que apresentam uma prevalência alta da patologia.

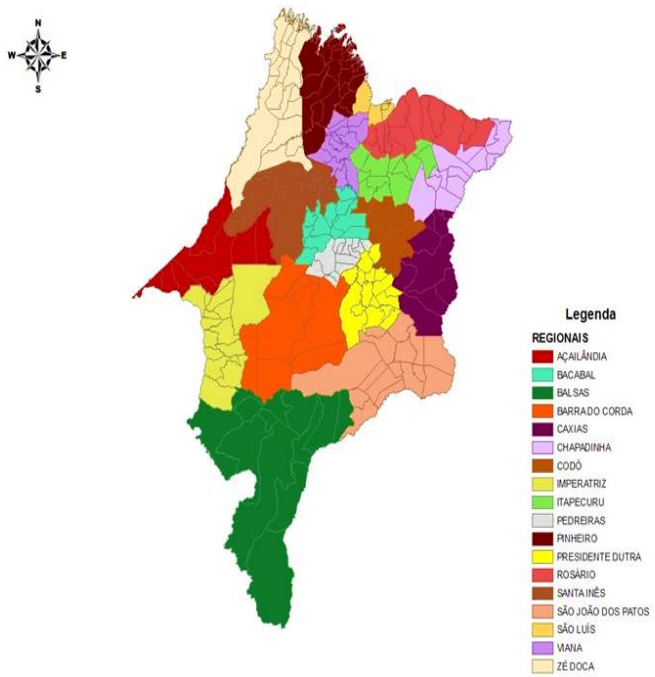
O estado do Maranhão é o segundo maior estado do Nordeste em extensão territorial e possui a segunda maior extensão litorânea do Brasil. O Maranhão está dividido em 217 munícipios e segundo o IBGE tem uma população estimada de 6.574.789 pessoas (Censo 2010) e sua densidade demográfica de 19,81 hab/km^2. Além disso, está em último lugar em comparação as outras Unidades da Federação quando o tema é rendimento nominal domiciliar per capita com o valor de R$ 636,00. As informações acima estão disponíveis em <https://cidades.ibge.gov.br/brasil/ma/panorama>. Com uma população estimada de 1.082.935 (Censo 2016) São Luís é o munícipio mais populoso do Maranhão, já em termos de extensão territorial o munícipio de Balsas está em primeiro lugar com 13.141,757 km^2.

A tuberculose é uma doença infecciosa e transmissível que afeta prioritariamente os pulmões, embora possa atingir outros órgãos e sistemas. A transmissão é feita através de gotículas eliminadas pela respiração, por espirros e pela tosse. Um dos principais sintomas da tuberculose é a presença de tosse por mais de três semanas com ou sem a presença de catarro que pode ser acompanhada ou não de febre ao final do dia. Por exemplo, no ano de 2019 foram registrados 2688 casos no estado do Maranhão segundo os dados do Ministério da Saúde/SVS - Sistema de Informação de Agravos de Notificações -

Sinan Net <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinannet/cnv/tubercma.def>.

O mapa do estado do Maranhão também pode ser dividido em unidades regionais de Saúde (ver Figura 5.1). Ao todo são 18 unidades regionais sediadas nos municípios de Rosário, Itapecuru, Chapadinha, Codó, Caxias, Presidente Dutra, Santa Inês, Zé Doca, Viana, Pinheiro, Bacabal, Pedreiras, Barra do Corda, Imperatriz, Açailândia, Balsas e São João dos Patos, bem como a sede, em São Luís, atendendo a 217 municípios maranhenses.

Figura 5.1: O mapa do estado do Maranhão dividido em unidades regionais de Saúde..



Fonte: Setor de Epidemiologia e Estatística (SEE) da Agência Estadual de Defesa Agropecuária (AGED) do Maranhão.

Primeiro iremos fazer um apanhado geral da ocorrência de tuberculose na população residente no estado do Maranhão. Se consideramos o período de 2001 a 2019 veremos que foram registrados 51.819 casos de tuberculose e 1.185 óbitos em decorrência da doença, que resulta em uma taxa anual de aproximadamente 63 óbitos/ano. Na Tabela 5.1 são mostrados em detalhes os números absolutos de ocorrência e óbitos de 2001 a 2019 no estado do Maranhão.

Tabela 5.1: Tuberculose - casos confirmados e óbito por tuberculose notificados no sistema de informação de agravos de notificação - Maranhão

Ano Diagnóstico	Casos confirmados	Óbito por tuberculose
2001	3.088	4
2002	3.204	2
2003	3.164	2
2004	3.180	2
2005	3.381	5
2006	3.056	32
2007	2.978	94
2008	2.628	68
2009	2.588	79
2010	2.518	74
2011	2.585	74
2012	2.280	109
2013	2.360	94
2014	2.168	71
2015	2.281	76
2016	2.487	84
2017	2.495	78
2018	2.690	115
2019	2.688	122
Total	51.819	1.185

Fonte: Ministério da Saúde/SVS - Sistema de Informação de Agravos de Notificação - Sinan Net.

Agora iremos apresentar uma análise espacial do tempo até o óbito por tuberculose no estado do Maranhão nos anos de 2011 a 2017. O conjunto de dados foi fornecido pelo **Prof. Dr. Max Sousa de Lima** da **Universidade Federal do Amazonas**. Foi realizada uma etapa de depuração dos dados com o objetivo de eliminar pacientes com dados incompletos ou inconsistentes. Além disso, só consideramos pacientes que foram a óbito em decorrência da tuberculose. A partir das variáveis presentes no banco de dados, foram criadas duas nova variáveis:

- `agrav`: que assume o valor 1, se o paciente possuía algum agravamento provocado por doenças pré-existentes e 0, caso contrário;
- `hiv`: que assume o valor 1, se o paciente havia testado positivo para o HIV-AIDS

e 0, caso contrário;

A variável resposta de interesse é:

- `tempo`: que foi o tempo do diagnóstico da doença até o óbito do paciente, medido em anos.

Depois da análise preliminar, o total de pacientes que ficou no conjunto de dados foi 419. A partir disso, considerando o modelo de regressão BP, estudamos de forma preliminar a relação da variável `tempo` com diferentes covariáveis como, por exemplo, `sexo`, `idade`, `escolaridade`, `local de moradia (urbano, zural)`, `raça`, `agravamento` e `hiv`. Desta forma, após um ajuste preliminar, chegamos a conclusão que apenas as variáveis agravamento e `hiv` foram estatisticamente significativas para o modelo e serão consideradas para o restante do estudo.

Na Tabela 5.2, apresentamos alguns resultados descritivos do tempo até o óbito com relação as variáveis `agravamento` e `hiv`. Nota-se que pacientes com agravamento apresentaram tempo de vida menores que pacientes sem doenças pré-existentes (aproximadamente de 11% menor). Além disso, se o paciente era portador de HIV, o tempo de vida era cerca de 46% menor do que quem era HIV negativo. Por fim, tivermos que o tempo médio de vida dos pacientes foi de 0,3 anos (desvio-padrão de 0,7 anos) após o diagnóstico da doença.

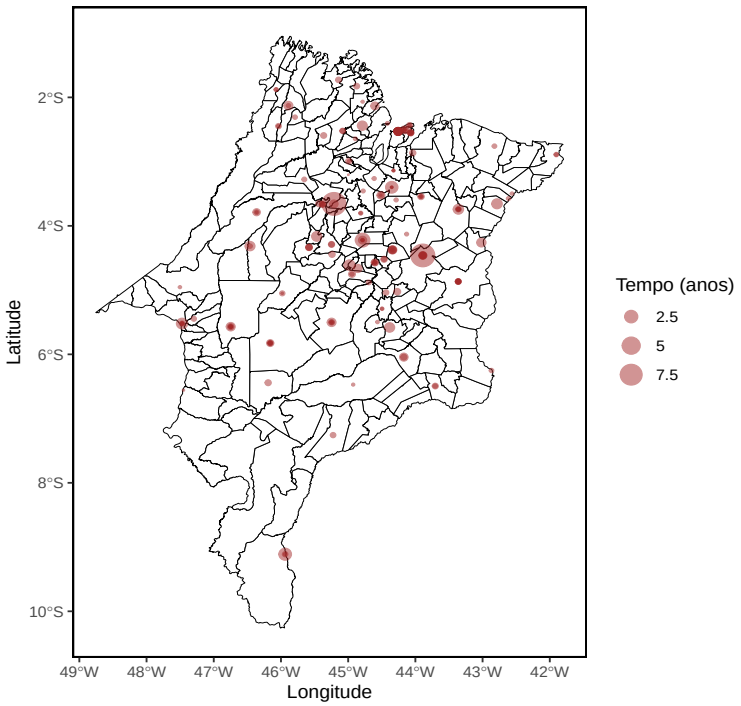
Tabela 5.2: Médias e Desvios-Padrão do tempo de vida, por status de agravamento e HIV, para 419 pacientes diagnosticados com tuberculose.

Condição	Status	n	Média	Desvio-Padrão
Agravamento	Sim	98	0,274	0,925
	Não	321	0,307	0,617
HIV	Sim	100	0,180	0,209
	Não	319	0,337	0,790

Fonte: Próprio autor

Na Figura 5.2, apresentamos a localização geográfica dos pacientes em que os pontos indicam um tempo de vida maior ou menor de acordo com o diâmetro do mesmo. De modo geral, nota-se uma maior concentração de pacientes na parte central do estado do Maranhão.

Figura 5.2: Distribuição dos pacientes de acordo com o tempo de vida após o diagnóstico de tuberculose.



Fonte: Próprio autor

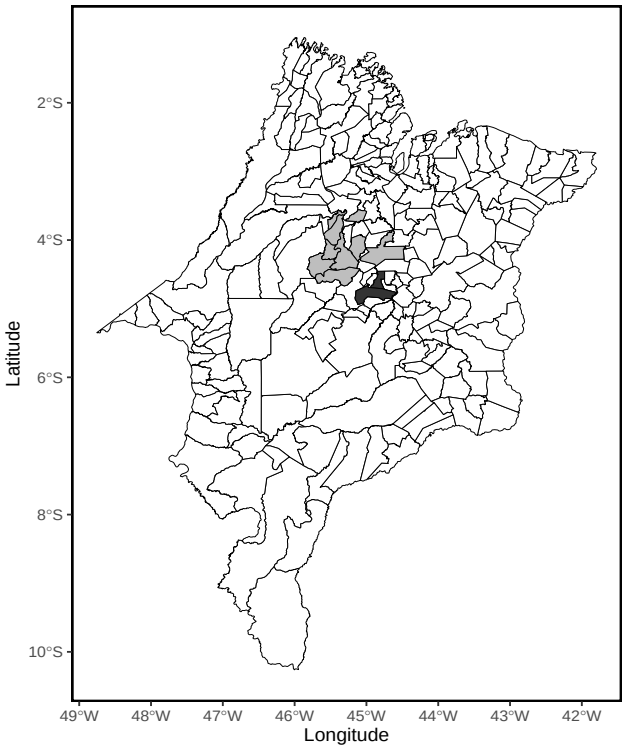
Aplicamos o modelo BP-SCAN considerando a possível relação entre o tempo até o óbito do paciente com a presença ou não de agravamentos e ter testado positivo ou não para HIV. Consideramos um modelo de regressão BP com a seguinte relação funcional

$$\log(\mu_i) = \beta_0 + \beta_1 \cdot \texttt{agrav} + \beta_2 \cdot \texttt{hiv} \quad \sigma_i = \sigma \quad (i = 1, \ldots, 419).$$

Os valores das estimativas foram: $\hat{\beta}_0 = -1,162$, $\hat{\beta}_1 = -0,349$, $\hat{\beta}_2 = -0,240$ e $\hat{\sigma} = 2,177$. Enquanto isso, o valor da estatística de teste foi 9,99. Para 100 réplicas de *bootstrap* obtivemos um p-valor de $0,0099$ com o *cluster* estimado $\hat{Z}$ formado pelos pacientes residentes em Pindaré-Mirim, Brejo de Areia, Igarapé do Meio, Bacabal, Santa Inês, Lago dos Rodrigues, Paulo Ramos, Igarapé Grande, Olho d'água das Cunhãs, Vitorino Freire, Altamira do Maranhão e Poção de Pedras (ver Figura 5.3).

O parâmetro de clusterização foi estimado em $\hat{\tau} = 0,826$, que pode ser interpretado como uma razão de chances $\exp\{0,826\} = 2,285$, significando que pacientes das áreas

Figura 5.3: Zonas identificadas como vulneráveis pelo modelo BP-SCAN.



Fonte: Próprio autor

destacadas possuem duas vezes mais chances de irem a óbito quando diagnosticados com tuberculose.

Capítulo 6

Considerações Finais

Neste estudo, propomos as estatísticas *scan* espacial NI-SCAN, GA-SCAN, BSR-SCAN, BP-SCAN e WE-SCAN, baseadas em modelos de regressão com variáveis respostas assimétricas. Sendo assim, propomos esses modelos para detectar *clusters* geográficos em dados contínuos distribuídos no intervalo $(0,\infty)$.

Quanto à simulação, o modelo NI-SCAN apresentou pior desempenho em relação aos demais métodos utilizados na simulação, pois os valores das medidas de poder do teste, sensibilidade e valor predito positivo foram menores em comparação aos valores dessas medidas para os demais modelos. De modo geral, o modelo BP-SCAN apresentou um melhor desempenho para detectar o *cluster*. Além disso, os resultados obtidos nas simulações nesta pesquisa, principalmente para $\sigma = 20$, corrobora com o estudo de (BHATT; TIMARI, 2014), em que estudos de simulação revelaram que o método da estatística *scan* espacial apresentou bom desempenho para diferentes distribuições de sobrevivência.

Adicionalmente, notamos que à medida que os parâmetros τ e σ aumentaram, a sensibilidade e valor predito positivo também aumentaram para todos os modelos utilizados na simulação. Além disso, para valores de $\sigma = \{0, 5, 2\}$ todos os modelos apresentaram poder de teste baixo. Verificou-se que o poder do teste foi maior no cenário com $\sigma = 20$ para todos os modelos com exceção do modelo NI-SCAN que apresentou poder de teste baixo em todos os cenários da simulação.

Conclui-se o trabalho com uma aplicação mostrando que a importância de se

estudar a teoria da estatística *scan* espacial está no fato que ela tem larga aplicabilidade em diversas áreas como: saúde (mapeamento de doenças e epidemiologia espacial), ambiental (monitoramento de problemas ambientais), análise criminal, genética de populações, astronomia, entre outras áreas. Sendo assim, a estatística *scan* espacial tem notável importância e amplo uso teórico e prático. Além disso, como possíveis trabalhos futuros podemos destacar: (i) Realizar a comparação dos modelos considerando outras estatísticas de teste como a estatística gradiente; (ii) Propor estatísticas *scan* considerando as versões com zero ajustados dos modelo aqui estudados; (iii) Propor estatísticas *scan* baseadas nos modelo estudados nesta dissertação considerando a presença de censura.

Apêndice A

Demonstração da Estatística de Teste do Modelo NI-SCAN

Sob o modelo NI-SCAN$(\mu_l, \sigma, \tau), l = 1, \ldots, L$.

$$f(y_l|\mu_l, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2 y_l^3}} \exp\left\{-\frac{1}{2\mu_l^2\sigma^2 y_l}(y_l - \mu_l)^2\right\},$$

Se $\hat{\tau} > 0$, então podemos escrever (3.2) como

$$\begin{aligned}
\hat{\Lambda}_Z^{NI} &= \left\{\log \ell_Z(\widehat{\boldsymbol{\theta}}_1, \hat{\tau}; \boldsymbol{y}) - \log \ell_0(\widehat{\boldsymbol{\theta}}_0, 0; \boldsymbol{y})\right\} \\
&= \sum_{s_l \notin Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) + \sum_{s_l \in Z} \log(\hat{\mu}_{Zl}, \hat{\sigma}) - \sum_{s_l \notin Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) - \sum_{s_l \in Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) \\
&= \sum_{s_l \in Z} \{\log(\hat{\mu}_{Zl}, \hat{\sigma}) - \log(\hat{\mu}_{0l}, \hat{\sigma})\},
\end{aligned}$$

denote $\hat{\Lambda}_l^{NI} = \{\log(\hat{\mu}_{Zl}, \hat{\sigma}) - \log(\hat{\mu}_{0l}, \hat{\sigma})\}$. Logo,

$$\begin{aligned}
\hat{\Lambda}_l^{NI} =& \log\left[\frac{1}{\sqrt{2\pi\hat{\sigma}^2 y_l^3}} \exp\left\{-\frac{1}{2\hat{\mu}_{Zl}^2\hat{\sigma}^2 y_l}(y_l - \hat{\mu}_{Zl})^2\right\}\right] \\
&- \log\left[\frac{1}{\sqrt{2\pi\hat{\sigma}^2 y_l^3}} \exp\left\{-\frac{1}{2\hat{\mu}_{0l}^2\hat{\sigma}^2 y_l}(y_l - \hat{\mu}_{0l})^2\right\}\right] \\
=& -\frac{1}{2}\log(2\pi\hat{\sigma}^2 y_l^3) - \frac{(y_l - \hat{\mu}_{Zl})^2}{2\hat{\mu}_{Zl}^2\hat{\sigma}^2 y_l} + \frac{1}{2}\log(2\pi\hat{\sigma}^2 y_l^3) + \frac{(y_l - \hat{\mu}_{0l})^2}{2\hat{\mu}_{0l}^2\hat{\sigma}^2 y_l}
\end{aligned}$$

$$
\begin{aligned}
&= - \frac{(y_l - \hat{\mu}_{Zl})^2}{2\hat{\mu}_{Zl}^2\hat{\sigma}^2 y_l} + \frac{(y_l - \hat{\mu}_{0l})^2}{2\hat{\mu}_{0l}^2\hat{\sigma}^2 y_l} \\
&= - \frac{y_l^2}{2\hat{\mu}_{Zl}^2\hat{\sigma}^2 y_l} + \frac{2y_l\hat{\mu}_{Zl}}{2\hat{\mu}_{Zl}^2\hat{\sigma}^2 y_l} - \frac{\hat{\mu}_{Zl}^2}{2\hat{\mu}_{Zl}^2\hat{\sigma}^2 y_l} + \frac{y_l^2}{2\hat{\mu}_{0l}^2\hat{\sigma}^2 y_l} - \frac{2y_l\hat{\mu}_{0l}}{2\hat{\mu}_{0l}^2\hat{\sigma}^2 y_l} + \frac{\hat{\mu}_{0l}^2}{2\hat{\mu}_{0l}^2\hat{\sigma}^2 y_l} \\
&= - \frac{y_l}{2\hat{\mu}_{Zl}^2\hat{\sigma}^2} + \frac{1}{\hat{\mu}_{Zl}\hat{\sigma}^2} - \frac{1}{2\hat{\sigma}^2 y_l} + \frac{y_l}{2\hat{\mu}_{0l}^2\hat{\sigma}^2} - \frac{1}{\hat{\mu}_{0l}\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^2 y_l} \\
&= - \frac{y_l}{2\hat{\mu}_{Zl}^2\hat{\sigma}^2} + \frac{1}{\hat{\mu}_{Zl}\hat{\sigma}^2} + \frac{y_l}{2\hat{\mu}_{0l}^2\hat{\sigma}^2} - \frac{1}{\hat{\mu}_{0l}\hat{\sigma}^2}.
\end{aligned}
$$

Substituindo $\exp\{\hat{\tau}\} = \dfrac{\hat{\mu}_{Zl}}{\hat{\mu}_{0l}}$, nota-se que $\hat{\mu}_{Zl} = \exp\{\hat{\tau}\} \times \hat{\mu}_{0l}$ e $\hat{\mu}_{Zl}^2 = \exp\{2\hat{\tau}\} \times \hat{\mu}_{0l}^2$ consequentemente,

$$
\begin{aligned}
\hat{\Lambda}_l^{NI} =& - \frac{y_l}{2e^{2\hat{\tau}}\hat{\mu}_{0l}^2\hat{\sigma}^2} + \frac{1}{e^{\hat{\tau}}\hat{\mu}_{0l}\hat{\sigma}^2} + \frac{y_l}{2\hat{\sigma}^2\hat{\mu}_{0l}^2} - \frac{1}{\hat{\mu}_{0l}\hat{\sigma}^2} \\
=& \frac{e^{-\hat{\tau}}}{\hat{\mu}_{0l}\hat{\sigma}^2} - \frac{y_l e^{-2\hat{\tau}}}{2\hat{\mu}_{0l}^2\hat{\sigma}^2} - \frac{1}{\hat{\mu}_{0l}\hat{\sigma}^2} + \frac{y_l}{2\hat{\sigma}^2\hat{\mu}_{0l}^2} \\
=& \frac{1}{\hat{\sigma}^2\widehat{\mu}_{0l}} \left(\mathrm{e}^{-\hat{\tau}} - 1\right) \left\{1 - \frac{y_l}{2\widehat{\mu}_{0l}} \left(\mathrm{e}^{-\hat{\tau}} + 1\right)\right\}.
\end{aligned}
$$

Portanto a prova está completa para a estatística de teste do modelo NI-SCAN$(\mu_l, \sigma, \tau), l = 1, \ldots, L$.

Apêndice B

Demonstração da Estatística de Teste do Modelo GA-SCAN

Sob o modelo GA-SCAN(μ_l, σ, τ),

$$f(y_l|\mu_l, \sigma) = \frac{y_l^{(1/\sigma^2-1)} \exp[-y_l/(\sigma^2\mu_l)]}{(\sigma^2\mu_l)^{(1/\sigma^2)}\Gamma(1/\sigma^2)}.$$

Quando $\hat{\tau} > 0$, podemos reescrever (3.2) como

$$\begin{aligned}
\widehat{\Lambda}_Z^{GA} &= \left\{\log \ell_Z(\widehat{\boldsymbol{\theta}}_1, \hat{\tau}; \boldsymbol{y}) - \log \ell_0(\widehat{\boldsymbol{\theta}}_0, 0; \boldsymbol{y})\right\} \\
&= \sum_{s_l \notin Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) + \sum_{s_l \in Z} \log(\hat{\mu}_{Zl}, \hat{\sigma}) - \sum_{s_l \notin Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) - \sum_{s_l \in Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) \\
&= \sum_{s_l \in Z} \{\log(\hat{\mu}_{Zl}, \hat{\sigma}) - \log(\hat{\mu}_{0l}, \hat{\sigma})\}.
\end{aligned}$$

Seja $\hat{\Lambda}_l^{GA} = \{\log(\hat{\mu}_{Zl}, \hat{\sigma}) - \log(\hat{\mu}_{0l}, \hat{\sigma})\}$ então,

$$\begin{aligned}
\hat{\Lambda}_l^{GA} &= \log\left[\frac{y_l^{(1/\hat{\sigma}^2-1)} \exp[-y_l/(\hat{\sigma}^2\hat{\mu}_{Zl})]}{(\hat{\sigma}^2\hat{\mu}_{Zl})^{(1/\hat{\sigma}^2)}\Gamma(1/\hat{\sigma}^2)}\right] - \log\left[\frac{y_l^{(1/\hat{\sigma}^2-1)} \exp[-y_l/(\hat{\sigma}^2\hat{\mu}_{0l})]}{(\hat{\sigma}^2\hat{\mu}_{0l})^{(1/\hat{\sigma}^2)}\Gamma(1/\hat{\sigma}^2)}\right] \\
&= \left(\frac{1}{\hat{\sigma}^2} - 1\right)\log(y_l) - \frac{y_l}{\hat{\sigma}^2\hat{\mu}_{Zl}} - \frac{1}{\hat{\sigma}^2}\log(\hat{\sigma}^2\hat{\mu}_{Zl}) - \log\Gamma\left(\frac{1}{\hat{\sigma}^2}\right) \\
&\quad - \left(\frac{1}{\hat{\sigma}^2} - 1\right)\log(y_l) + \frac{y_l}{\hat{\sigma}^2\hat{\mu}_{0l}} + \frac{1}{\hat{\sigma}^2}\log(\hat{\sigma}^2\hat{\mu}_{0l}) + \log\Gamma\left(\frac{1}{\hat{\sigma}^2}\right) \\
&= -\frac{y_l}{\hat{\sigma}^2\hat{\mu}_{Zl}} + \frac{y_l}{\hat{\sigma}^2\hat{\mu}_{0l}} - \frac{1}{\hat{\sigma}^2}\log(\hat{\sigma}^2\hat{\mu}_{Zl}) + \frac{1}{\hat{\sigma}^2}\log(\hat{\sigma}^2\hat{\mu}_{0l})
\end{aligned}$$

$$
\begin{aligned}
=&\frac{y_l}{\hat{\sigma}^2}\left[\frac{1}{\hat{\mu}_{0l}}-\frac{1}{\hat{\mu}_{Zl}}\right]-\frac{1}{\hat{\sigma}^2}\left[\log(\hat{\sigma}^2\hat{\mu}_{Zl})-\log(\hat{\sigma}^2\hat{\mu}_{0l})\right]\\
=&\frac{y_l}{\hat{\sigma}^2}\left[\frac{1}{\hat{\mu}_{0l}}-\frac{1}{\hat{\mu}_{Zl}}\right]-\frac{1}{\hat{\sigma}^2}\log\left[\frac{\hat{\sigma}^2\hat{\mu}_{Zl}}{\hat{\sigma}^2\hat{\mu}_{0l}}\right].
\end{aligned}
$$

Substituindo $\exp\{\hat{\tau}\}=\dfrac{\hat{\mu}_{Zl}}{\hat{\mu}_{0l}}$ tem-se,

$$
\begin{aligned}
\hat{\Lambda}_l^{GA}=&\frac{y_l}{\hat{\sigma}^2}\left[\frac{1}{\hat{\mu}_{0l}}-\frac{1}{\hat{\mu}_{Zl}}\right]-\frac{1}{\hat{\sigma}^2}\log\left[e^{\hat{\tau}}\right]\\
=&\frac{y_l}{\hat{\sigma}^2}\left[\frac{1}{\hat{\mu}_{0l}}-\frac{1}{\hat{\mu}_{Zl}}\right]-\frac{\hat{\tau}}{\hat{\sigma}^2}\\
=&\frac{y_l}{\hat{\sigma}^2}\left[\frac{1}{\hat{\mu}_{0l}}-\frac{1}{\hat{\mu}_{0l}e^{\hat{\tau}}}\right]-\frac{\hat{\tau}}{\hat{\sigma}^2}\\
=&\frac{y_l}{\hat{\sigma}^2}\left[\frac{1}{\hat{\mu}_{0l}}-\frac{e^{-\hat{\tau}}}{\hat{\mu}_{0l}}\right]-\frac{\hat{\tau}}{\hat{\sigma}^2}\\
=&\frac{y_l(1-e^{-\hat{\tau}})}{\hat{\sigma}^2\hat{\mu}_{0l}}-\frac{\hat{\tau}}{\hat{\sigma}^2}\\
=&\frac{1}{\hat{\sigma}^2}\left[\frac{y_l(1-e^{-\hat{\tau}})}{\hat{\mu}_{0l}}-\hat{\tau}\right].
\end{aligned}
$$

Portanto a demonstração está completa para a estatística de teste do modelo GA-SCAN(μ_l,σ,τ).

Apêndice C

Demonstração da Estatística de Teste do Modelo WE-SCAN

Considere $\hat{\tau} > 0$, então escrevendo (3.2) de outra forma, ou seja,

$$\begin{aligned}\hat{\Lambda}_Z^{WE} &= \left\{\log \ell_Z(\widehat{\boldsymbol{\theta}}_1, \hat{\tau}; \boldsymbol{y}) - \log \ell_0(\widehat{\boldsymbol{\theta}}_0, 0; \boldsymbol{y})\right\} \\ &= \sum_{s_l \notin Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) + \sum_{s_l \in Z} \log(\hat{\mu}_{Zl}, \hat{\sigma}) - \sum_{s_l \notin Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) - \sum_{s_l \in Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) \\ &= \sum_{s_l \in Z} \{\log(\hat{\mu}_{Zl}, \hat{\sigma}) - \log(\hat{\mu}_{0l}, \hat{\sigma})\},\end{aligned}$$

em que sob o modelo WE-SCAN(μ_l, σ, τ),

$$f(y_l|\mu_l, \sigma) = \frac{\sigma}{\left(\frac{\mu_l}{\Gamma(1+\frac{1}{\sigma})}\right)^{\sigma}} y_l^{(\sigma-1)} exp\left\{-\left(\frac{y_l \Gamma(1+\frac{1}{\sigma})}{\mu_l}\right)^{\sigma}\right\}, \quad y_l, \mu_l, \sigma \in (0, \infty),$$

considere $\hat{\Lambda}_l^{WE} = \{\log(\hat{\mu}_{Zl}, \hat{\sigma}) - \log(\hat{\mu}_{0l}, \hat{\sigma})\}$ dessa forma,

$$\begin{aligned}\hat{\Lambda}_l^{WE} = &\log\left[\frac{\hat{\sigma}}{\left(\frac{\hat{\mu}_{Zl}}{\Gamma(1+\frac{1}{\hat{\sigma}})}\right)^{\hat{\sigma}}} y_l^{(\hat{\sigma}-1)} exp\left\{-\left(\frac{y_l \Gamma(1+\frac{1}{\hat{\sigma}})}{\hat{\mu}_{Zl}}\right)^{\hat{\sigma}}\right\}\right] \\ &- \log\left[\frac{\hat{\sigma}}{\left(\frac{\hat{\mu}_{0l}}{\Gamma(1+\frac{1}{\hat{\sigma}})}\right)^{\hat{\sigma}}} y_l^{(\hat{\sigma}-1)} exp\left\{-\left(\frac{y_l \Gamma(1+\frac{1}{\hat{\sigma}})}{\hat{\mu}_{0l}}\right)^{\hat{\sigma}}\right\}\right]\end{aligned}$$

$$
\begin{aligned}
&= \log(\hat{\sigma}) - \hat{\sigma}\log\left(\frac{\hat{\mu}_{Zl}}{\Gamma(1+\frac{1}{\hat{\sigma}})}\right) + (\hat{\sigma}-1)\log(y_l) - \left(\frac{y_l\Gamma(1+\frac{1}{\hat{\sigma}})}{\hat{\mu}_{Zl}}\right)^{\hat{\sigma}} \\
&\quad - \log(\hat{\sigma}) + \hat{\sigma}\log\left(\frac{\hat{\mu}_{0l}}{\Gamma(1+\frac{1}{\hat{\sigma}})}\right) - (\hat{\sigma}-1)\log(y_l) + \left(\frac{y_l\Gamma(1+\frac{1}{\hat{\sigma}})}{\hat{\mu}_{0l}}\right)^{\hat{\sigma}} \\
&= \left(\frac{y_l\Gamma(1+\frac{1}{\hat{\sigma}})}{\hat{\mu}_{0l}}\right)^{\hat{\sigma}} - \left(\frac{y_l\Gamma(1+\frac{1}{\hat{\sigma}})}{\hat{\mu}_{Zl}\times\frac{\hat{\mu}_{0l}}{\hat{\mu}_{0l}}}\right)^{\hat{\sigma}} - \hat{\sigma}\left[\log\left(\frac{\hat{\mu}_{Zl}}{\Gamma(1+\frac{1}{\hat{\sigma}})}\right) - \log\left(\frac{\hat{\mu}_{0l}}{\Gamma(1+\frac{1}{\hat{\sigma}})}\right)\right] \\
&= \left(\frac{y_l\Gamma(1+\frac{1}{\hat{\sigma}})}{\hat{\mu}_{0l}}\right)^{\hat{\sigma}} - \left(\frac{y_l\Gamma(1+\frac{1}{\hat{\sigma}})e^{-\hat{\tau}}}{\hat{\mu}_{0l}}\right)^{\hat{\sigma}} - \hat{\sigma}\log\left[\frac{\frac{\hat{\mu}_{Zl}}{\Gamma(1+\frac{1}{\hat{\sigma}})}}{\frac{\hat{\mu}_{0l}}{\Gamma(1+\frac{1}{\hat{\sigma}})}}\right] \\
&= \left(\frac{y_l\Gamma(1+\frac{1}{\hat{\sigma}})}{\hat{\mu}_{0l}}\right)^{\hat{\sigma}}(1-e^{-\hat{\tau}\hat{\sigma}}) - \hat{\sigma}\log\left[\frac{\hat{\mu}_{Zl}}{\Gamma(1+\frac{1}{\hat{\sigma}})}\times\frac{\Gamma(1+\frac{1}{\hat{\sigma}})}{\hat{\mu}_{0l}}\right],
\end{aligned}
$$

substituindo $\exp\{\hat{\tau}\} = \dfrac{\hat{\mu}_{Zl}}{\hat{\mu}_{0l}}$ tem -se que,

$$
\begin{aligned}
\hat{\Lambda}_l^{WE} &= \left(\frac{y_l\Gamma(1+\frac{1}{\hat{\sigma}})}{\hat{\mu}_{0l}}\right)^{\hat{\sigma}}(1-e^{-\hat{\tau}\hat{\sigma}}) - \hat{\sigma}\log\left(\exp\{\hat{\tau}\}\right) \\
&= \left(\frac{y_l\Gamma(1+\frac{1}{\hat{\sigma}})}{\hat{\mu}_{0l}}\right)^{\hat{\sigma}}(1-e^{-\hat{\tau}\hat{\sigma}}) - \hat{\sigma}\hat{\tau}.
\end{aligned}
$$

Logo a prova está completa para a estatística de teste do modelo WE-SCAN(μ_l, σ, τ).

Apêndice D

Demonstração da Estatística de Teste do Modelo BSR-SCAN

Sob o modelo BSR-SCAN(μ_l, σ, τ),

$$f(y_l|\mu_l,\sigma) = \frac{\exp(\sigma/2)\sqrt{\sigma+1}}{4\sqrt{\pi\mu_l}\, y_l^{3/2}}\left[y_l + \frac{\sigma\mu_l}{\sigma+1}\right]\exp\left(-\frac{\sigma}{4}\left[\frac{\{\sigma+1\}y_l}{\sigma\mu_l} + \frac{\sigma\mu_l}{\{\sigma+1\}y_l}\right]\right),$$

quando $\hat{\tau} > 0$ e reescrevendo (3.2) como

$$\begin{aligned}
\widehat{\Lambda}_Z^{BSR} &= \left\{\log \ell_Z(\widehat{\boldsymbol{\theta}}_1, \hat{\tau}; \boldsymbol{y}) - \log \ell_0(\widehat{\boldsymbol{\theta}}_0, 0; \boldsymbol{y})\right\} \\
&= \sum_{s_l \notin Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) + \sum_{s_l \in Z} \log(\hat{\mu}_{Zl}, \hat{\sigma}) - \sum_{s_l \notin Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) - \sum_{s_l \in Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) \\
&= \sum_{s_l \in Z} \{\log(\hat{\mu}_{Zl}, \hat{\sigma}) - \log(\hat{\mu}_{0l}, \hat{\sigma})\}.
\end{aligned}$$

Denote $\hat{\Lambda}_l^{BSR} = \{\log(\hat{\mu}_{Zl}, \hat{\sigma}) - \log(\hat{\mu}_{0l}, \hat{\sigma})\}$ sendo assim,

$$\begin{aligned}
\hat{\Lambda}_l^{BSR} =& \log\left[\frac{\exp(\hat{\sigma}/2)\sqrt{\hat{\sigma}+1}}{4\sqrt{\pi\hat{\mu}_{Zl}}\, y_l^{3/2}}\left[y_l + \frac{\hat{\sigma}\hat{\mu}_{Zl}}{\hat{\sigma}+1}\right]\exp\left(-\frac{\hat{\sigma}}{4}\left[\frac{\{\hat{\sigma}+1\}y_l}{\hat{\sigma}\hat{\mu}_{Zl}} + \frac{\hat{\sigma}\hat{\mu}_{Zl}}{\{\hat{\sigma}+1\}y_l}\right]\right)\right] \\
&- \log\left[\frac{\exp(\hat{\sigma}/2)\sqrt{\hat{\sigma}+1}}{4\sqrt{\pi\hat{\mu}_{0l}}\, y_l^{3/2}}\left[y_l + \frac{\hat{\sigma}\hat{\mu}_{0l}}{\hat{\sigma}+1}\right]\exp\left(-\frac{\hat{\sigma}}{4}\left[\frac{\{\hat{\sigma}+1\}y_l}{\hat{\sigma}\hat{\mu}_{0l}} + \frac{\hat{\sigma}\hat{\mu}_{0l}}{\{\hat{\sigma}+1\}y_l}\right]\right)\right] \\
=& \frac{\hat{\sigma}}{2} + \log\left[\sqrt{\hat{\sigma}+1}\right] - \log\left[4\sqrt{\pi\hat{\mu}_{Zl}}\right] - \frac{3}{2}\log(y_l) + \log\left[y_l + \frac{\hat{\sigma}\hat{\mu}_{Zl}}{\hat{\sigma}+1}\right]
\end{aligned}$$

$$
\begin{aligned}
&- \frac{\hat{\sigma}}{4}\left[\frac{\{\hat{\sigma}+1\}y_l}{\hat{\sigma}\hat{\mu}_{Zl}} + \frac{\hat{\sigma}\hat{\mu}_{Zl}}{\{\hat{\sigma}+1\}y_l}\right] - \frac{\hat{\sigma}}{2} - \log\left[\sqrt{\hat{\sigma}+1}\right] + \log\left[4\sqrt{\pi\hat{\mu}_{0l}}\right] \\
&+ \frac{3}{2}\log(y_l) - \log\left[y_l + \frac{\hat{\sigma}\hat{\mu}_{0l}}{\hat{\sigma}+1}\right] + \frac{\hat{\sigma}}{4}\left[\frac{\{\hat{\sigma}+1\}y_l}{\hat{\sigma}\hat{\mu}_{0l}} + \frac{\hat{\sigma}\hat{\mu}_{0l}}{\{\hat{\sigma}+1\}y_l}\right] \\
=&\log\left[y_l + \frac{\hat{\sigma}\hat{\mu}_{Zl}}{\hat{\sigma}+1}\right] - \log\left[y_l + \frac{\hat{\sigma}\hat{\mu}_{0l}}{\hat{\sigma}+1}\right] + \log\left[4\sqrt{\pi\hat{\mu}_{0l}}\right] - \log\left[4\sqrt{\pi\hat{\mu}_{Zl}}\right] \\
&- \frac{\hat{\sigma}}{4}\left[\frac{\{\hat{\sigma}+1\}y_l}{\hat{\sigma}\hat{\mu}_{Zl}} + \frac{\hat{\sigma}\hat{\mu}_{Zl}}{\{\hat{\sigma}+1\}y_l}\right] + \frac{\hat{\sigma}}{4}\left[\frac{\{\hat{\sigma}+1\}y_l}{\hat{\sigma}\hat{\mu}_{0l}} + \frac{\hat{\sigma}\hat{\mu}_{0l}}{\{\hat{\sigma}+1\}y_l}\right] \\
=&\log\left[\frac{\frac{\hat{\sigma}y_l + y_l + \hat{\sigma}\hat{\mu}_{Zl}}{\hat{\sigma}+1}}{\frac{\hat{\sigma}y_l + y_l + \hat{\sigma}\hat{\mu}_{0l}}{\hat{\sigma}+1}}\right] + \log\left[\frac{4\sqrt{\pi\hat{\mu}_{0l}}}{4\sqrt{\pi\hat{\mu}_{Zl}}}\right] - \frac{\hat{\sigma}\{\hat{\sigma}+1\}y_l}{4\hat{\sigma}\hat{\mu}_{Zl}} \\
&- \frac{\hat{\sigma}^2\hat{\mu}_{Zl}}{4\{\hat{\sigma}+1\}y_l} + \frac{\hat{\sigma}\{\hat{\sigma}+1\}y_l}{4\hat{\sigma}\hat{\mu}_{0l}} + \frac{\hat{\sigma}^2\hat{\mu}_{0l}}{4\{\hat{\sigma}+1\}y_l}.
\end{aligned}
$$

Considere a seguinte substituição $\exp\{\hat{\tau}\} = \dfrac{\hat{\mu}_{Zl}}{\hat{\mu}_{0l}}$, em que $\hat{\mu}_{Zl} = \exp\{\hat{\tau}\} \times \hat{\mu}_{0l}$ logo,

$$
\begin{aligned}
\hat{\Lambda}_l^{BSR} =&\log\left[\frac{\hat{\sigma}y_l + y_l + \hat{\sigma}\hat{\mu}_{Zl}}{\hat{\sigma}y_l + y_l + \hat{\sigma}\hat{\mu}_{0l}}\right] + \log\left(\sqrt{\frac{\hat{\mu}_{0l}}{\hat{\mu}_{Zl}}}\right) - \frac{\{\hat{\sigma}+1\}y_l}{4\hat{\mu}_{0l}e^{\hat{\tau}}} \\
&+ \frac{\{\hat{\sigma}+1\}y_l}{4\hat{\mu}_{0l}} - \frac{\hat{\sigma}^2 e^{\hat{\tau}}\hat{\mu}_{0l}}{4\{\hat{\sigma}+1\}y_l} + \frac{\hat{\sigma}^2\hat{\mu}_{0l}}{4\{\hat{\sigma}+1\}y_l} \\
=&\log\left[\frac{\hat{\sigma}y_l + y_l + \hat{\sigma}\hat{\mu}_{Zl}}{\hat{\sigma}y_l + y_l + \hat{\sigma}\hat{\mu}_{0l}}\right] + \frac{1}{2}\log\left(\frac{\hat{\mu}_{0l}}{\hat{\mu}_{0l}e^{\hat{\tau}}}\right) + \frac{y_l\{\hat{\sigma}+1\}}{4\hat{\mu}_{0l}}(1 - e^{-\hat{\tau}}) \\
&+ \frac{\hat{\sigma}^2\hat{\mu}_{0l}}{4\{\hat{\sigma}+1\}y_l}(1 - e^{\hat{\tau}}) \\
=&-\frac{\hat{\tau}}{2} + \log\left[\frac{\hat{\sigma}y_l + y_l + \hat{\sigma}\hat{\mu}_{Zl}}{\hat{\sigma}y_l + y_l + \hat{\sigma}\hat{\mu}_{0l}}\right] + \frac{y_l\{\hat{\sigma}+1\}}{4\hat{\mu}_{0l}}(1 - e^{-\hat{\tau}}) + \frac{\hat{\sigma}^2\hat{\mu}_{0l}}{4\{\hat{\sigma}+1\}y_l}(1 - e^{\hat{\tau}}).
\end{aligned}
$$

Então está completa a demonstração para a estatística de teste do modelo BSR-SCAN(μ_l, σ, τ).

Apêndice E

Demonstração da Estatística de Teste do Modelo BP-SCAN

Sob o modelo BP-SCAN(μ_l, σ, τ),

$$f(y_l|\mu_l, \sigma) = \frac{y_l^{\mu_l(\sigma+1)-1}(1+y_l)^{-[\mu_l(\sigma+1)+\sigma+2]}}{B(\mu_l(\sigma+1),(\sigma+2))},$$

seja $\hat{\tau} > 0$, então podemos reescrever (3.2) da seguinte maneira,

$$\begin{aligned}
\hat{\Lambda}_Z^{BP} &= \left\{\log \ell_Z(\widehat{\boldsymbol{\theta}}_1, \hat{\tau}; \boldsymbol{y}) - \log \ell_0(\widehat{\boldsymbol{\theta}}_0, 0; \boldsymbol{y})\right\} \\
&= \sum_{s_l \notin Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) + \sum_{s_l \in Z} \log(\hat{\mu}_{Zl}, \hat{\sigma}) - \sum_{s_l \notin Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) - \sum_{s_l \in Z} \log(\hat{\mu}_{0l}, \hat{\sigma}) \\
&= \sum_{s_l \in Z} \{\log(\hat{\mu}_{Zl}, \hat{\sigma}) - \log(\hat{\mu}_{0l}, \hat{\sigma})\},
\end{aligned}$$

denote $\hat{\Lambda}_l^{BP} = \{\log(\hat{\mu}_{Zl}, \hat{\sigma}) - \log(\hat{\mu}_{0l}, \hat{\sigma})\}$ logo,

$$\begin{aligned}
\hat{\Lambda}_l^{BP} =& \log\left[\frac{y_l^{\hat{\mu}_{Zl}(\hat{\sigma}+1)-1}(1+y_l)^{-[\hat{\mu}_{Zl}(\hat{\sigma}+1)+\hat{\sigma}+2]}}{B(\hat{\mu}_{Zl}(\hat{\sigma}+1),(\hat{\sigma}+2))}\right] - \log\left[\frac{y_l^{\hat{\mu}_{0l}(\hat{\sigma}+1)-1}(1+y_l)^{-[\hat{\mu}_{0l}(\hat{\sigma}+1)+\hat{\sigma}+2]}}{B(\hat{\mu}_{0l}(\hat{\sigma}+1),(\hat{\sigma}+2))}\right] \\
=& [\hat{\mu}_{Zl}(\hat{\sigma}+1)-1]\log(y_l) - [\hat{\mu}_{Zl}(\hat{\sigma}+1)+(\hat{\sigma}+2)]\log(1+y_l) \\
& - \log\left(B(\hat{\mu}_{Zl}(\hat{\sigma}+1),(\hat{\sigma}+2))\right) - [\hat{\mu}_{0l}(\hat{\sigma}+1)-1]\log(y_l) \\
& + [\hat{\mu}_{0l}(\hat{\sigma}+1)+(\hat{\sigma}+2)]\log(1+y_l) + \log\left(B(\hat{\mu}_{0l}(\hat{\sigma}+1),(\hat{\sigma}+2))\right)
\end{aligned}$$

$$
\begin{aligned}
=&[\hat{\mu}_{Zl}(\hat{\sigma}+1)]\log(y_l) - \log(y_l) - [\hat{\mu}_{Zl}(\hat{\sigma}+1)]\log(1+y_l) - (\hat{\sigma}+2)\log(1+y_l) \\
&- \log\left(B(\hat{\mu}_{Zl}(\hat{\sigma}+1),(\hat{\sigma}+2))\right) - [\hat{\mu}_{0l}(\hat{\sigma}+1)]\log(y_l) + \log(y_l) \\
&+ [\hat{\mu}_{0l}(\hat{\sigma}+1)]\log(1+y_l) + (\hat{\sigma}+2)\log(1+y_l) + \log\left(B(\hat{\mu}_{0l}(\hat{\sigma}+1),(\hat{\sigma}+2))\right) \\
=&(\hat{\sigma}+1)\log(y_l)[\hat{\mu}_{Zl} - \hat{\mu}_{0l}] - (\hat{\sigma}+1)\log(1+y_l)[\hat{\mu}_{Zl} - \hat{\mu}_{0l}] \\
&+ \log\left(\frac{B(\hat{\mu}_{0l}(\hat{\sigma}+1),(\hat{\sigma}+2))}{B(\hat{\mu}_{Zl}(\hat{\sigma}+1),(\hat{\sigma}+2))}\right) \\
=&(\hat{\sigma}+1)[\hat{\mu}_{Zl} - \hat{\mu}_{0l}][\log(y_l) - \log(1+y_l)] + \log\left(\frac{B(\hat{\mu}_{0l}(\hat{\sigma}+1),(\hat{\sigma}+2))}{B(\hat{\mu}_{Zl}(\hat{\sigma}+1),(\hat{\sigma}+2))}\right) \\
=&(\hat{\sigma}+1)[\hat{\mu}_{Zl} - \hat{\mu}_{0l}]\log\left(\frac{y_l}{1+y_l}\right) + \log\left(\frac{B(\hat{\mu}_{0l}(\hat{\sigma}+1),(\hat{\sigma}+2))}{B(\hat{\mu}_{Zl}(\hat{\sigma}+1),(\hat{\sigma}+2))}\right).
\end{aligned}
$$

Substituindo $\exp\{\hat{\tau}\} = \dfrac{\hat{\mu}_{Zl}}{\hat{\mu}_{0l}}$ então.

$$
\begin{aligned}
\hat{\Lambda}_l^{BP} =&(\hat{\sigma}+1)[\hat{\mu}_{0l}e^{\hat{\tau}} - \hat{\mu}_{0l}]\log\left(\frac{y_l}{1+y_l}\right) + \log\left(\frac{B(\hat{\mu}_{0l}(\hat{\sigma}+1),(\hat{\sigma}+2))}{B(\hat{\mu}_{Zl}(\hat{\sigma}+1),(\hat{\sigma}+2))}\right) \\
=&\hat{\mu}_{0l}(\hat{\sigma}+1)[e^{\hat{\tau}} - 1]\log\left(\frac{y_l}{1+y_l}\right) + \log\left(\frac{B(\hat{\mu}_{0l}(\hat{\sigma}+1),(\hat{\sigma}+2))}{B(\hat{\mu}_{Zl}(\hat{\sigma}+1),(\hat{\sigma}+2))}\right).
\end{aligned}
$$

Sendo assim, está completa a demonstração para a estatística de teste do modelo BP-SCAN(μ_l, σ, τ).

Referências Bibliográficas

BESCULIDES, M. et al. Evaluation of school absenteeism data for early outbreak detection, new york city. *BMC Public Health*, v. 5, n. 1, p. 105, 2005. Disponível em: <https://doi.org/10.1186/1471-2458-5-105>.

BHATT, V.; TIMARI, N. Scan statistics for survival data based on weibull distribution. *Statistics in Medicine*, v. 33, n. 11, p. 1867–1876, May 2014.

BIRNBAUM, Z.; SAUNDERS, S. A new family of life distributions. *Journal of Applied Probability*, v. 6, n. 2, p. 319–327, August 1969.

BOURGUIGNON, M.; SANTOS-NETO, M.; CASTRO, M. de. A new regression model for positive data. 2018. Disponível em: <http://arxiv.org/pdf/1804.07734v1>.

BUSE, A. The likelihood ratio, wald, and lagrange multiplier tests: An expository note. *The American Statistician*, [American Statistical Association, Taylor & Francis, Ltd.], v. 36, n. 3, p. 153–157, 1982. ISSN 00031305. Disponível em: <www.jstor.org/stable/2683166>.

CANÇADO, A. F.; FERNANDES, L. B.; SILVA, C. Q. da. A bayesian spatial scan statistic for zero-inflated count data. *Spatial Statistics*, v. 20, n. Supplement C, p. 57–75, 2017. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2211675317300374>.

CASELLA, G.; BERGER, R. L. *Statistical Inference*. [S.l.]: California, Duxbury., 2002.

DUCZMAL, L.; KULLDORFF, M.; HUANG, L. Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 15, n. 2, p. 428–442, 2006. Disponível em: <https://doi.org/10.1198/106186006X112396>.

DWASS, M. On the distribution of ranks and of certain rank order statistics. *Ann. Math. Statist.*, The Institute of Mathematical Statistics, v. 28, n. 2, p. 424–431, 06 1957. Disponível em: <https://doi.org/10.1214/aoms/1177706970>.

EFRON, B. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, The Institute of Mathematical Statistics, v. 7, n. 1, p. 1–26, 01 1979. Disponível em: <https://doi.org/10.1214/aos/1176344552>.

ELIAS, J. et al. Spatiotemporal analysis of invasive meningococcal disease, germany. *Emerging Infectious Diseases*, Centers for Disease Control and Prevention, v. 12, n. 11, p. 1689–1695, 11 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3372358/>.

GRATZER, G. *Math into LaTeX : an introduction to LaTeX and AMS-LaTex*. Boston: Birkhä user, 1996. ISBN 978-0817638054.

HUANG, L.; KULLDORFF, M.; GREGORIO, D. A spatial scan statistics for survival data. *Biometrics*, v. 63, n. 1, p. 109–118, March 2007.

HUANG, L. et al. Covariate adjusted weighted spatial scan statistics with applications to study geographic clustering of obesity and lung cancer mortality in the united states. *Statist in Medicine*, v. 29, n. 23, p. 2410–2422, October 2010.

JOHANSEN, A. M. et al. Monte carlo methods. In: *International Encyclopedia of Education (Third Edition)*. Oxford: Elsevier, 2010. p. 296–303. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780080448947015438>.

JUNG, I.; KULLDORFF, M.; KLASSEN, A. C. A spatial scan statistic for ordinal data. *Statistics in Medicine*, John Wiley & Sons, Ltd., v. 26, n. 7, p. 1594–1607, 2007. Disponível em: <http://dx.doi.org/10.1002/sim.2607>.

JUNG, I.; KULLDORFF, M.; RICHARD, O. A spatial scan statistic for multinomial data. *Statist in Medicine*, v. 29, n. 18, p. 1910–1918, August 2010.

KEEPING, E. *Introduction to Statistical Inference*. New York: Van Nostrand, 1962.

KULLDORFF, M. A spatial scan statistic. *Comm. Statist. Theory Meth.*, v. 26, n. 6, p. 1481–1496, June 1997.

KULLDORFF, M.; HUANG, L.; KONTY, K. A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics*, BioMed Central, v. 8, p. 58–58, 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2772848/>.

KULLDORFF, M.; TANGO, T.; PARK, P. J. Power comparisons for disease clustering tests. *Computational Statistics and Data Analysis*, v. 42, n. 4, p. 665–684, 2003.

LEÃO, J. et al. Birnbaum–saunders frailty regression models: Diagnostics and application to medical data. *Biometrical Journal*, Version of Record online, January 2017.

LEIVA, V. *The Birnbaum-Saunders distribution*. 1st. ed. [S.l.]: Academic Press, 2016.

LEIVA, V. et al. A criterion for environmental assessment using birnbaum-saunders attribute control charts. *Environmetrics*, v. 26, n. 7, p. 463–476, November 2015.

LIMA, M. S. *Método adaptativo para detecção de clusters no espaço-tempo.* Tese (Doutorado) — Universidade Federal de Minas Gerais., 2011.

LIMA, M. S. de et al. A spatial scan statistic for beta regression. *Spatial Statistics*, v. 18, n. Part B, p. 444–454, November 2016.

MCDONALD, J. Some generalized functions for the size distribution of income. *Econometrica*, [Wiley, Econometric Society], v. 52, p. 647–663, 1984. Disponível em: <http://www.jstor.org/stable/1913469>.

MINAMISAVA, R. et al. Spatial clusters of violent deaths in a newly urbanized region of brazil: highlighting the social disparities. *International Journal of Health Geographics*, v. 8, n. 1, p. 66, 2009. Disponível em: <https://doi.org/10.1186/1476-072X-8-66>.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, v. 54, p. 507–554, 2005. ISSN 0035-9254.

SANTOS-NETO, M. *Estimação e modelagem com a distribuição Birnbaum-Saunders: uma nova reparametrização.* Dissertação (Mestrado) — Universidade Federal de Pernambuco., 2010.

SANTOS-NETO, M. *RBS: Regression models for Birnbaum-Saunders distributions.* [S.l.], 2020. R package version 1.0-3.

SANTOS-NETO, M. et al. On new parameterizations of the birnbaum-saunders distribution. *Pak. J. Statist.*, v. 28, n. 1, p. 1–26, January/February 2012.

SANTOS-NETO, M. et al. Reparameterized birnbaum-saunders regression models with varying precision. *Electronic Journal of Statistics*, v. 10, n. 2, p. 2825–2855, September 2016.

SATO, S.; INOUE, J. Inverse gaussian distribution and its application. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, Wiley Subscription Services, Inc., A Wiley Company, v. 77, n. 1, p. 32–42, 1994. Disponível em: <http://dx.doi.org/10.1002/ecjc.4430770104>.

TEAM, R. D. C. *R: A Language and Environment for Statistical Computing.* Vienna, Austria, 2020.

THOM, H. *A note on the gamma distribution.* Statistical Laboratory, 1947. Manuscript.

WIJNGAARD, C. C. van den et al. Syndromic surveillance for local outbreaks of lower-respiratory infections: Would it work? *PLOS ONE*, Public Library of Science, v. 5, n. 4, p. e10406–, 04 2010. Disponível em: <https://doi.org/10.1371/journal.pone.0010406>.